IN THE UNITED STATES DISTRICT COURT
FOR THE SOUTHERN DISTRICT OF TEXAS
HOUSTON DIVISION

DWIGHT BAZILE, *et al.*,	§	
	§	
Plaintiffs,	§	
	§	
VS.	§	CIVIL ACTION NO. H-08-2404
	§	
CITY OF HOUSTON,	§	
	§	
Defendant.	§	

MEMORANDUM AND OPINION

This Title VII disparate-impact suit challenges the City of Houston's system for promoting firefighters to the positions of captain and senior captain. Historically, the City has promoted firefighters based on their years of service with the Houston Fire Department ("HFD") and their scores on a multiple-choice exam. The format and general content of that exam are set out in the Texas Local Government Code ("TLGC") and in the collective bargaining agreement ("CBA") between the City and the firefighters' union, the Houston Professional Fire Fighters Association ("HPFFA"). Seven black firefighters sued the City, alleging that the promotional exams for the captain and senior-captain positions were racially discriminatory, in violation of the Fourteenth Amendment, 42 U.S.C. § 1981, and Title VII, 42 U.S.C. § 2000e-2. After mediation in February and March 2010, the City and the seven firefighters reached a settlement that included a proposed consent decree. The decree would require the City to implement changes to the captain and senior-captain promotion exams in two phases. The first phase required minor changes to the November 2010 captain exam. The second phase required more significant changes to the May 2011 senior-captain exam that would apply to future captain and senior-captain exams. The HPFFA intervened

in the lawsuit and objected because the proposed consent decree changed the exams in ways inconsistent with the TLGC and the CBA. The HPFFA contended that the City and the plaintiffs had not shown discrimination that would permit this court to approve the consent decree.

This court bifurcated the proceedings to resolve the HPFFA's objections. The first stage addressed the narrow set of changes proposed for the November 2010 captain exam. The second stage addressed the broader changes proposed for subsequent captain and senior-captain exams. In the first stage, this court, with the HPFFA's agreement, approved changes to the November exam. This opinion addresses the HPFFA's objections to the proposed changes to the future captain and senior-captain exams.

The HPFFA vigorously objects that the proposed changes involve "a far-reaching and wholesale restructuring of the entire promotional process that goes beyond anything plaintiffs have even alleged in this lawsuit" and "bypass both the long-established protections of state law and the union's protected role in being the sole, collective voice for the city's firefighters." (Docket Entry No. 89, at 8). The City and the seven individual plaintiffs acknowledge that the changes are far-reaching but argue that they are needed to comply with federal antidiscrimination law. They argue that the current exam system is not "job related for the position[s] in question and consistent with business necessity" as required by 42 U.S.C. § 2000e-2(k)(1)(A)(i).

This court held an evidentiary hearing to consider the proposed changes to the May 2011 senior-captain exam and to future exams. The summary of the evidence shows the welter of expert opinions the parties presented on whether the existing format and content of the City's promotion exams for the captain and senior-captain positions have a disparate impact on African-American candidates; whether the existing exams are reliable and valid measures of the knowledge and

qualities relevant to the promotion decisions; whether the existing exams are reliable and valid ways to compare candidates; and whether the proposed changes to the exams will provide reliable and valid exams and address disparate impact. The experts' testimony and submissions left the court with a sense of disquiet about the opinions expressed. The science of testing to measure and compare promotion-worthiness is admittedly imperfect. The expert witnesses, particularly for the City, acknowledged some errors and some incomplete aspects of their work in designing and administering the promotion exams. At best, all the witnesses' opinions amount to uncertain efforts to gauge how well different exam approaches measure, compare, and predict job performance. The analytical steps required by the applicable legal standards must be approached with a recognition of the limits of the expert testimony.

At the same time, courts clearly lack expertise in the area of testing validity. "'The study of employment testing, although it has necessarily been adopted by the law as a result of Title VII and related statutes, is not primarily a legal subject.' Because of the substantive difficulty of test validation, courts must take into account the expertise of test validation professionals." *Gulino v. N.Y. State Educ. Dep't*, 460 F.3d 361, 383 (2d Cir. 2006) (quoting *Guardians Ass'n of N.Y.C. Police Dep't, Inc. v. Civil Serv. Comm'n of City of N.Y.*, 630 F.2d 79, 89 (2d Cir. 1980)). The combination of the lack of judicial expertise in this area and the limits of the expertise of those who do have training and experience support a cautious and careful judicial approach.

Based on the parties' filings, the evidence, and the applicable law, this court finds that the City and the seven individual plaintiffs have shown that the captain and senior-captain exams violate Title VII. But this court also finds that some of the changes in the proposed consent decree violate the CBA and TLGC and that the City and the plaintiffs have not shown that all these changes are

necessary to comply with Title VII. Based on these findings and conclusions, the proposed consent decree is accepted in part and denied in part. The use of situational-judgment questions and an assessment center are justified by the record evidence and are job-related and consistent with business necessity. But other parts of the proposed modified consent decree violate the TLGC and CBA, and the City and the plaintiffs have not shown that they are tailored to respond to the disparate impact alleged. Using the parties' descriptions of the proposed changes, the provisions that violate the TLGC and CBA without the necessary justification in the record, and which this court does not accept, are as follows:

2. Job-Knowledge Written Test
   - Pass/fail test (test designer to determine cut-off score)
   - No rank-order list
   - Test designer may elect not to use any written job-knowledge cognitive test

3. Scenario-Based Computer-Objective Test
   - Rank-order list from which all intended promotions to assessment center will be made[1]

4. Assessment Center
   - Rank-order list
   - Sliding bands based on test accuracy as determined by consultant
   - Fire Chief will document reasons for the selection of each candidate within bands
   - No Rule of Three

(Docket Entry No. 69-2, at 29).

The reasons for finding these aspects of the proposed changes to the promotion examinations

[1] This provision was modified after this court issued its ruling on the November 2010 captain exam. The modifications included changing the provision stating that the rank-order list resulting from the scenario-based computer-objective test would be used to generate a rank-order list "from which 150-200% of intended promotions to assessment center" to a provision stating that the test would generate a rank-order list "from which all of intended promotions to Assessment Center" would be made. (*Compare* Docket Entry No. 69-2, at 29, *with* Docket Entry No. 86-1, at 3).

invalid, and the remaining aspects supported by the record and the applicable law, are explained below. This opinion first describes the promotion system in place before any changes; reviews the expert and other evidence relevant to assessing disparate impact; and analyzes whether, under the applicable law, the proposed settlement is tailored to remedying the disparate impact that is shown. A hearing is set for **February 21, 2012, at 1:30 p.m.** to address the issues that remain to be resolved and a timetable for doing so.

Finally, because the terminology used by the HFD and the industrial psychologists who served as experts in this case produced a number of acronyms and abbreviations, a list of the most commonly used is attached to this Memorandum and Opinion.

I. Background

A. The Houston Fire Department

The HFD has approximately 4,000 employees involved in firefighting. Ninety percent are in the Emergency Operations Division ("EOD"). Half of the EOD employees are at the "firefighter" level and perform "task-level jobs" such as retrieving and using fire hoses. The next rank above firefighter is "engineer operator" ("EO"). In addition to performing firefighters' tasks, EOs drive fire trucks and HFD ambulances and operate ladders and pumps. Firefighters outnumber EOs two-to-one. (Evidentiary Hr'g Tr. 115, Docket Entry No. 130).

Captains are ranked immediately above EOs. HFD captains are the "first line of supervisor position[s] in the fire department." A captain supervises the operation of fire engines, which are smaller fire trucks that carry hoses and pump water. Each HFD fire station has at least one fire engine and one captain. A captain supervises an EO and two firefighters assigned to an engine.

When a captain misses a day of work, an EO may "ride up" and perform the absent captain's job duties.

Senior captains are ranked immediately above captains. A senior captain supervises the operations of "ladder trucks," which are large fire trucks with aerial ladders. Only half of the City's fire stations have a ladder truck with a senior captain in addition to a fire engine and captain. A senior captain may supervise up to eight firefighters, including EOs. When a senior captain misses a day of work, a captain may "ride up."

During a fire emergency, a district chief—ranked above senior captain—is responsible for developing the firefighting strategy. The district chief may decide, for example, whether firefighters will enter a burning building and address a fire directly or instead contain it by protecting adjacent buildings. Senior captains may participate in the strategy development, but the district chief bears ultimate responsibility. Once a strategy is set, the senior captain and captain are responsible for implementing it. Usually a senior captain and the ladder-truck crew are responsible for forcible entries into a building to ventilate it, for attempting rescues, and for creating ways for other firefighters to enter. The captain and the fire-engine crew are usually responsible for locating, confining, and extinguishing fires. (*Id.* at 116–20).

To summarize the promotional system that is discussed in detail below, promotion from EO to captain and from captain to senior captain depends largely on a candidate's score on a multiple-choice test. Any person meeting the experience requirement can take the test. An EO can apply for captain after four years in the fire department. A captain can apply for senior captain after two additional years of service as a captain. TEX. LOC. GOV'T CODE § 143.028(a). A candidate's

length of service with the HFD will add some points to the test score, but the test score largely determines promotion.

The City makes promotion decisions based on a rank-order list of the candidates' test points added to their length-of-service points. For each captain or senior-captain position available during the three years after the exam, the top three candidates' names and scores are submitted to the HFD fire chief. The presumption is that the fire chief will select the candidate with the highest test score. If the fire chief selects the second or third highest scoring candidate, the chief must explain his reasons in writing. If a candidate is not selected for promotion within the three-year period, the candidate must retake the exam. These promotional procedures for the captain and senior-captain positions are based on the TLGC and the CBA.

1. The Texas Local Government Code

The City of Houston adopted the Fire Fighter and Police Civil Service Act ("CSA"), codified as Chapter 143 of the TLGC, on January 31, 1948.[2] The CSA's "fundamental principle" is ensuring that public-service appointments and promotions are made "according to merit and fitness, ascertained by competitive examinations." *Klinger v. City of San Angelo*, 902 S.W.2d 669, 671 (Tex. App.—Austin 1995, writ denied). The Texas legislature passed the CSA "to secure efficient fire and police departments composed of capable personnel who are free from political influence." TEX. LOC. GOV'T CODE § 143.001(a). The TLGC requires a test-based promotional system for firefighters. Section 143.021(c) states that positions within fire departments must be filled "from an eligibility list that results from an examination held in accordance with [the CSA]." The TLGC

[2] *See* Charter of the City of Houston, Appendix B (2006), *available at* http://library.municode.com/HTML/10123/level2/AP_APXBREORADEL.html; TEX. LOC. GOV'T CODE § 143.002 (providing that the CSA applies only to municipalities who adopt it by election); TEX. LOC. GOV'T CODE § 143.004 (describing adoption procedures).

contains detailed rules describing the eligibility list, exam, and procedure for selecting firefighters for promotion.

The promotional process begins when a city posts notice of an upcoming examination. Municipalities like the City of Houston, with populations greater than 1.5 million, must post notice in plain view on a bulletin board located in City Hall's main lobby and in the Firefighters' and Police Officers' Civil Service Commission office by the 90th day before the date a promotional exam is scheduled. This 90-day notice must show the positions to be filled and the date, time, and place of the exam. TEX. LOC. GOV'T CODE § 143.107(a). The 90-day notice must also list the sources from which the exam questions are taken. *Id.* § 143.029(a). By the 30th day before the date a promotional exam is scheduled, another notice must be posted in the same locations. *Id.* § 143.107(b). The 30-day notice must state the number of newly created positions and may "include the name of each source used for the examination, the number of questions taken from each source, and the chapter used in each source." *Id.* § 143.029(c).

The TLGC requires that the test be in writing and forbids tests that "in any part consist of an oral interview." *Id.* § 143.032(c). The questions must "test the knowledge of the eligible promotional candidates about information and facts." *Id.* § 143.032(d). The information-and-fact questions "must" be based on:

> (1) the duties of the position for which the examination is held;
>
> (2) material that is of reasonably current publication and that has been made reasonably available to each member of the fire or police department involved in the examination; and
>
> (3) any study course given by the departmental schools of instruction.

Id. The questions must also be taken from the sources identified in the posted notices. *Id.* § 143.032(e). Finally, the "examination questions must be prepared and composed so that the grading of the examination can be promptly completed immediately after the examination is over." *Id.* § 143.032(f).

The exam grade determines whether the candidate will be placed on a promotion-eligibility list. Grading begins as soon as an individual candidate completes the exam. The candidate may remain present during the grading. *Id.* § 143.033(a). The multiple-choice exam score is based on a maximum grade of 100 points and is determined by the correctness of the answers to the questions. *Id.* § 143.033(c). Each candidate also receives one point for each year of seniority, with a maximum of 10 points. *Id.* § 143.033(b). In municipalities like Houston, a candidate must score at least 70 points on the exam to be eligible for promotion. *Id.* § 143.108(a).

All scores must be posted within 24 hours of the exam. *Id.* § 143.033(d). Each candidate may see the answers, grading, and source materials after the exam and can appeal a score within 5 days. *Id.* § 143.034(a). The City has 60 days to decide the appeal. *Id.* § 143.1015(a). A candidate who appeals is entitled to a hearing. *Id.* § 143.1015(b).

Once the scores are finalized, all candidates who pass are listed in rank order on a promotion-eligibility list. *See id.* § 143.021(c); *id.* § 143.108(f). When vacancies occur, the names of the three persons with the highest scores for the position are certified and provided to the head of the department with the vacancy. *Id.* § 143.036(b). This is known as the "Rule of Three." The TLGC provides that "[u]nless the department head has a valid reason" for not doing so, "the department head shall appoint the eligible promotional candidate having the highest grade on the eligibility list." *Id.* § 143.036(f). If the candidate with the highest grade is not selected, the

department head must personally discuss the reason with that candidate and file a written explanation. *Id.*

2. The Collective Bargaining Agreement

Texas law establishes firefighters' right to collective bargaining. TEX. LOC. GOV'T CODE § 174.002(b) ("The policy of this state is that fire fighters and police officers, like employees in the private sector, should have the right to organize for collective bargaining, as collective bargaining is a fair and practical method for determining compensation and other conditions of employment. Denying fire fighters and police officers the right to organize and bargain collectively would lead to strife and unrest, consequently injuring the health, safety, and welfare of the public."); *id.* § 143.204(a) (stating that a firefighter association submitting a petition signed by the majority of the paid firefighters in the municipality "may be recognized . . . as the sole and exclusive bargaining agent for all of the covered fire fighters"). The HPFFA is the sole and exclusive bargaining agent for the City's firefighters.

The TLGC allows the City and the HPFFA to enter into a written agreement binding when ratified by both. *Id.* § 143.206(a). Such an agreement can supersede the TLGC's provisions "concerning wages, salaries, rates of pay, hours of work, and other terms and conditions of employment to the extent of any conflict with the [written agreement]." *Id.* § 143.207(a). The agreement "preempts all contrary local ordinances, executive orders, legislation, or rules adopted by the state." *Id.* § 143.207(b).

The 2009-2010 CBA between the City of Houston and the HPFFA made few departures from the TLGC's exam provisions. Like the TLGC, the CBA required a grade of at least 70% for promotion eligibility. The CBA specified that the test must consist of "not less than 100 and not

more than 150 questions." (Docket Entry No. 69-6, at 20). Unlike the TLGC, the CBA allowed only a .5-point increase in the score for each year of service, with a maximum of 10 points. The CBA also allowed a .5-point increase for each year of service for certain ranks. For example, an engineer or operator applying to be a captain is awarded .5 points for each year of service as an engineer. (*Id.*). Aside from these changes, the 2009-2010 CBA provided that the TLGC "remain[s] in full force in the same manner as on the date [the CBA] became effective." (*Id.* at 13).

B. Title VII

"Congress enacted Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e et seq., to assure equality of employment opportunities by eliminating those practices and devices that discriminate on the basis of race, color, religion, sex, or national origin." *Alexander v. Gardner-Denver Co.*, 415 U.S. 36, 44 (1974). Title VII's prohibitions include using "a particular employment practice that causes a disparate impact on the basis of race, color, religion, sex, or national origin" unless the employment practice "is job related for the position in question and consistent with business necessity." 42 U.S.C. § 2000e-2(k)(1)(A)(i). The plaintiffs alleged that the City's promotional procedures for captain and senior captain violated Title VII's disparate-impact provision.

"Congress intended voluntary compliance to be the preferred means of achieving the objectives of Title VII." *Local No. 93, Int'l Ass'n of Firefighters, AFL-CIO v. City of Cleveland*, 478 U.S. 501, 515 (1986). To help employers comply with Title VII, Congress authorized the Equal Employment Opportunity Commission ("EEOC") to issue compliance guidelines (the "Guidelines"). The Guidelines "are not administrative regulations promulgated pursuant to formal procedures established by the Congress. But . . . they do constitute '(t)he administrative interpretation of the

Act by the enforcing agency,' and consequently they are 'entitled to great deference.'" *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 431 (1975) (citing *Griggs v. Duke Power Co.*, 401 U.S. 424, 433–34 (1971)).

The Guidelines require employers who make promotional decisions based on test scores to maintain records of tests and test results. 29 C.F.R. § 1607.4(A). The Guidelines' rule of thumb for determining disparate impact is the "4/5 Rule." Under this Rule:

> A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of disparate impact.

Id. § 1607.4(D). There are exceptions to the 4/5 Rule. The Guidelines state:

> Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms or where a user's actions have discouraged applicants disproportionately on grounds of race, sex, or ethnic group. Greater differences in selection rate may not constitute adverse impact where the differences are based on small numbers and are not statistically significant, or where special recruiting or other programs cause the pool of minority or female candidates to be atypical of the normal pool of applicants from that group. Where the user's evidence concerning the impact of a selection procedure indicates adverse impact but is based upon numbers which are too small to be reliable, evidence concerning the impact of the procedure over a longer period of time and/or evidence concerning the impact which the selection procedure had when used in the same manner in similar circumstances elsewhere may be considered in determining adverse impact. Where the user has not maintained data on adverse impact as required by the documentation section of applicable guidelines, the Federal enforcement agencies may draw an inference of adverse impact of the selection process from the failure of the user to maintain such data, if the user has an underutilization of a group in the job category, as compared to the group's representation in the relevant labor market or, in the case of jobs filled from within, the applicable work force.

Id.

If analyzing an employer's test results under the 4/5 Rule shows "that the total selection process for a job has an adverse impact, the individual components of the selection process should be evaluated for adverse impact." *Id.* § 1607.4(C). The method for evaluating individual components is a "validity study." *Id.* § 1607.3(A). The Guidelines describe three types of validity studies: criterion-related-validity studies; content-validity studies; and construct-validity studies. *Id.* § 1607.5(A). A criterion-related-validity study analyzes whether test results correlate to "criteria that [are] predictive of job performance." Mark R. Bandsuch, *Ten Troubles with Title VII and Trait Discrimination Plus One Simple Solution (A Totality of the Circumstances Framework)*, 37 CAP. U. L. REV. 965, 1089 (2009). A content-validity study analyzes whether test results correlate to "the knowledge, skills, and abilities related to that job." *Id.* A construct-validity study examines whether test results correlate to "general characteristics important to job performance." *Id.*

The Guidelines also describe the evidence each type of validity study requires. A criterion-related-validity study requires "empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance." 29 C.F.R. § 1607.5(B). A content-validity study requires "data showing that the content of the selection procedure is representative of important aspects of performance on the job for which the candidates are to be evaluated." *Id.* A construct-validity study requires "data showing that the procedure measures the degree to which candidates have identifiable characteristics which have been determined to be important in successful performance in the job for which the candidates are to be evaluated." *Id.*

One court has summarized the content-validity and criterion-validity methods for evaluating a promotion or other employment test, as follows:

> [E]mployers can establish job-relatedness by one of three methods, including "content validity," which entails showing that the test measures the job or adequately reflects the skills or knowledge required by the job. A typing test for secretaries exemplifies this kind of approach. This method does not require empirical evidence, but instead "should consist of data showing that the content of the selection procedure is representative of important aspects of performance on the job." 29 C.F.R. § 1607.5(B). In contrast, the "criterion related" approach evaluates whether a test is adequately correlated with future job performance and is constructed to measure traits thought to be relevant to future job performance. An IQ test is a typical criterion-related method. Unlike content validity, this method requires "empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance." 29 C.F.R. § 1607.5(B).

Banos v. City of Chicago, 398 F.3d 889, 893 (7th Cir. 2005) (citations and internal quotations marks omitted).[3]

Before conducting a validity study, an employer should conduct a "job analysis." 29 C.F.R. § 1607.14(A). Each type of validity study requires a different type of job analysis. Criterion-related- validity studies require "reviewing job information to determine measures of work behavior(s) or performance that are relevant to the job or group of jobs in question." *Id.* § 1607.14(B)(2). "These measures or criteria are relevant to the extent that they represent critical or important job duties, work behaviors or work outcomes as developed from the review of job information"; "[b]ias should be considered."[4] *Id.* Content-validity studies should include "an

[3] 29 C.F.R. § 1607.14 describes the technical standards for each type of validity study in greater detail. These standards are discussed and applied below.

[4] In addition to work behaviors and performance, certain criteria—such as production rate, error rate, tardiness, absenteeism, and length of service—may be used without a full job analysis if the user can show the importance of the criteria to the particular employment context. The Guidelines state:

analysis of the important work behavior(s) required for successful performance and their relative importance and, if the behavior results in work product(s), an analysis of the work product(s). Any job analysis should focus on the work behavior(s) and the tasks associated with them." *Id.* § 1607.14(C)(2). Construct-validity studies "should show the work behavior(s) required for successful performance of the job, or the groups of jobs being studied, the critical or important work behavior(s) in the job or group of jobs being studied, and an identification of the construct(s) believed to underlie successful performance of these critical or important work behaviors in the job or jobs in question." *Id.* § 1607.14(D)(2). "Each construct should be named and defined, so as to distinguish it from other constructs." *Id.*

If one or more validity studies produces evidence "sufficient to warrant use of the procedure for the intended purpose under the standard of these guidelines," the promotional procedure is "properly validated." *Id.* § 1607.16(X). But if no validity study produces sufficient evidence, an employer "should initiate affirmative steps to remedy the situation." *Id.* § 1607.17(3). These steps, "which in design and execution may be race, color, sex, or ethnic 'conscious,' include, but are not limited to," the following:

> (a) The establishment of a long-term goal, and short-range, interim goals and timetables for the specific job classifications, all of which

---

> A standardized rating of overall work performance may be used where a study of the job shows that it is an appropriate criterion. Where performance in training is used as a criterion, success in training should be properly measured and the relevance of the training should be shown either through a comparison of the content of the training program with the critical or important work behavior(s) of the job(s), or through a demonstration of the relationship between measures of performance in training and measures of job performance. Measures of relative success in training include but are not limited to instructor evaluations, performance samples, or tests. Criterion measures consisting of paper and pencil tests will be closely reviewed for job relevance.

29 C.F.R. 1607.14(B)(3).

> should take into account the availability of basically qualified persons in the relevant job market;
>
> (b) A recruitment program designed to attract qualified members of the group in question;
>
> (c) A systematic effort to organize work and redesign jobs in ways that provide opportunities for persons lacking "journeyman" level knowledge or skills to enter and, with appropriate training, to progress in a career field;
>
> (d) Revamping selection instruments or procedures which have not yet been validated in order to reduce or eliminate exclusionary effects on particular groups in particular job classifications;
>
> (e) The initiation of measures designed to assure that members of the affected group who are qualified to perform the job are included within the pool of persons from which the selecting official makes the selection;
>
> (f) A systematic effort to provide career advancement training, both classroom and on-the-job, to employees locked into dead end jobs; and
>
> (g) The establishment of a system for regularly monitoring the effectiveness of the particular affirmative action program, and procedures for making timely adjustments in this program where effectiveness is not demonstrated.

Id.

C. The Procedural History of this Case

On August 4, 2008, seven firefighters sued the City of Houston, alleging that the 2006 captain and senior-captain exams had a discriminatory effect on their promotion opportunities, in violation of § 1981 and 42 U.S.C. § 2000e-2. The seven individual plaintiffs contended that the 2006 exams had a disparate impact on the promotion of black firefighters to captain and senior-captain positions compared to white firefighters. Four plaintiffs—Dwight Bazile, Johnny Garrett, Trevin Hines, and Mundo Olford—were lieutenants denied promotion to captain. Three

plaintiffs—George Runnels, Dwight Allen, and Thomas Ward—were captains denied promotion to senior captain. (Docket Entry No. 1).

The City and the plaintiffs settled. (Docket Entry No. 64). The HPFFA was not a party to the negotiations or settlement. The City agreed to promote Bazile, Olford, and Hines to captain; to promote Allen to senior captain; to allow Garrett to retire as a captain; and to allow Runnels and Ward to retire as senior captains. The City also agreed to pay each plaintiff backpay in amounts ranging from $376.80 to $23,075.46. (Docket Entry No. 69-2, at 2–8, 17–22, 26–27).

The settlement also contained a proposed consent decree to be submitted to the court for approval. (*Id.* at 9–10, 28–30). The decree required the City to implement changes to the captain and senior-captain exams in two phases. In the first phase, the City agreed to implement "modest" changes to the November 2010 captain exam. In the second phase, the City agreed to implement broader changes, beginning with the May 2011 senior-captain exam and applying to all future captain and senior-captain exams. The settlement agreement required the parties to give notice to the HPFFA of "this conceptual agreement" and to "meet in person or conference call [with the HPFFA] to explore potential adjustments of union suggestions prior to final settlement meeting." (*Id.* at 10). The settlement agreement also required approval by the Houston City Council and by this court.

The parties notified this court of the settlement and their intent to file the proposed consent decree. Before filing the decree, the parties moved to join the HPFFA to the suit because the proposed changes to the promotion exams conflicted with the TLGC and the CBA. (Docket Entry No. 69). The HPFFA moved to intervene and asked this court to bifurcate review of the proposed consent decree. The first step would be to consider the HPFFA's objections to the proposed changes

to the November 2010 captain exam. The second stage would be to consider the HPFFA's objections to the proposed changes to the subsequent senior-captain and later captain and senior-captain exams. This court granted the motion and entered a scheduling order. (Docket Entry Nos. 70 & 71).

1. The November 2010 Captain Exam

The HPFFA objected to certain proposed changes to the November 2010 captain exam. (Docket Entry No. 75). This court heard arguments and evidence on the objections on September 16, 2010. On the same date, and with the HPFFA's agreement, this court found that the existing captain exam disparately impacted black firefighters and entered an order allowing the City to implement the consent decree provisions changing the November 2010 captain exam. (Docket Entry Nos. 82 & 85). The proposed consent decree described those changes to the 2010 captain exam, as follows:

> Hybrid written examination
> Content-validated job-knowledge portion of exam
> - weighted to Houston Departmental material, plus
> - carefully selected directly relevant test material,
> - supported by job analysis and incumbent/supervisor feedback in direct interviews
> - job analysis
>
> Content-validated multiple-choice situational-judgment exam
> - designed by an industrial/organization psychologist
> - HFD departmental scenario based
> - zero, partial, and full-credit options
> - additional points to stay the same as under the Collective Bargaining Agreement in an effort to minimize adverse impact potential
> - rank order for selection process by Chief
> - Rule of Three for promotion

(Docket Entry No. 69-2, at 28–29).

The most significant change to the November 2010 captain exam was the inclusion of multiple-choice "situational-judgment" questions in addition to the "job-knowledge" questions used on previous exams. Situational-judgment questions present hypothetical situations encountered on the job and ask candidates how they would respond.[5] By contrast, the job-knowledge questions that made up the previous captain exams, mandated by the TLGC, *see* TEX. LOC. GOV'T CODE § 143.032(d), ask about facts related to the job, such as the content of applicable regulations or specific HFD policies and procedures.[6]

This court's order approved the inclusion of situational-judgment multiple-choice questions for the November 2010 captain exam based on a finding that "[t]he continued exclusive use of questions based on 'fact' and 'information' as stated in Local Government Code § 143.032(d) is likely to continue to result in adverse impact." (Docket Entry No. 85, at 2). The

---

[5] *See* Robert E. Ployhart & William I. MacKenzie, Jr., *Situational Judgment Tests: A Critical Review and Agenda for the Future*, *in* APA HANDBOOK OF INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY 237, 237 (S. Zedeck ed., 2010) ("Situational judgment tests . . . are measurement methods that present respondents with work-related situations and ask them how they would or should handle the situations.").

[6] An example of a job-knowledge question is: "Regarding command organization, the Incident Commander is responsible for [what] Level of the Command structure, which involves the overall Command of the incident." The answer choices are: (a) "Operational"; (b) "Tactical"; (c) "Task"; and (d) "Strategic." (Evidentiary Hr'g Ex. 32, Question 21).

A situational-judgment question begins by describing a scenario and then asks the candidate the most likely and least likely action he or she would take as fire captain from a list of options. An example is as follows:

> It is 0200 hours on a Sunday morning. Your company has . . . been dispatched to a structure fire . . . that was reported by a police officer. . . . You are first to arrive on the scene and give the following initial report: "Engine 522 is on the scene of a 200 X 300 foot building with a bow string truss roof. We have smoke and fire showing from roof area on the C side. Engine 522 is assuming . . . Command." You expect the next company to arrive in about four (4) minutes.

(*Id.*, Question 101). The candidate must choose from the following actions: (a) "[a]nnounce that this will be a defensive mode operation, order one of the firefighters to conduct a 360 of the structure, and order incoming trucks to spot the apparatus in preparation to fly pipe"; (b) "[o]rder the next-in Engine to establish a water supply and advance a line to the front door"; (c) "[a]nnounce that this will be an offensive operation, order the next-in engine to establish a water supply, and advance a 2 ½ [inch] line to the A side doors"; and (d) "[p]ass [c]ommand to the next-in company and investigate." (*Id.*).

situational-judgment questions included in the November 2010 captain exam were developed by industrial-psychology consultants selected by the City, the plaintiffs, and the HPFFA (the "consultants").[7] These consultants developed the questions by creating a job analysis for the captain position. *See* 29 C.F.R. § 1607.14(A) (describing a job analysis). To create the job analysis, the consultants interviewed "subject-matter experts" ("SMEs"); analyzed HFD materials related to the captain position, such as internal job descriptions and policies and procedures; and analyzed external source materials such as published industry standards and firefighting textbooks. The consultants interviewed both internal SMEs—incumbent HFD captains and their supervisors—and external SMEs—individuals with similar experience who did not work for HFD. Through the job analysis, the consultants identified the "knowledge, skills, abilities and other characteristics" ("KSAOs") required for successful performance in the captain position and designed the situational-judgment questions to measure the identified KSAOs.

This court also approved an additional consent-decree provision inconsistent with the TLGC and CBA. The TLGC and CBA allow a promotional candidate to be present while the candidate's exam is scored and require that scores be posted within 24 hours of the exam. TEX. LOC. GOV'T CODE § 143.033(a), (d). The consent decree required an "item analysis" of the score before it was finalized. Item analysis requires the consultants to aggregate data related to each question—or "item"—to eliminate questions that did not reliably measure an individual candidate's exam performance.[8] Because item analysis requires collecting data from all promotional candidates'

[7] The HPFFA's consultant was Dr. Winfred Arthur; the City's were Dr. David Morris and Dr. Kathleen Lundquist; the plaintiffs' was Dr. Kyle Brink. (Docket Entry No. 94-1, at 2).

[8] Two examples of item analyses are item-difficulty analysis and item-discrimination analysis. "Item difficulty is a statistic used in test analysis that indicates the percentage of applicants who answered an item correctly." (Evidentiary Hr'g Ex. 3, Dr. Brink Report, at 40). If too many or too few promotional candidates correctly answer the question, it may be eliminated (or all candidates may be given credit for answering the question correctly). The premise

exams and time to evaluate this data, the parties agreed to post the raw scores from the exam within 24 hours to meet the TLGC and the CBA requirements, but these raw scores would not be the final scores until the item analysis was completed. The promotional candidates would not remain throughout the item analysis.

Many of the TLGC and CBA requirements remained in place under the consent decree for the November 2010 captain exam. The consent decree still required job-knowledge questions. The exam that was administered contained 75 job-knowledge questions. The job analysis was included as a "source material" in the posted notices and made available before the exam. *See* TEX. LOC. GOV'T CODE § 143.029(a) (requiring posting of source material). The consent decree also required a rank order of the candidates based on their exam scores and required that promotions be made according to the Rule of Three set out in the TLGC. *See id.* § 143.036(b), (f) (describing the Rule of Three).

The City administered the captain exam on November 17, 2010. The consultants conducted an item analysis after the exam. A panel consisting of representatives for the City, the individual

---

for eliminating these questions is that if too high or too low a percentage of test-takers correctly answer the question, the question does not reliably demonstrate that one test-taker is more qualified than others. One study recommends that items with difficulties above .9 (90% of test-takers answered correctly) should be eliminated unless the test is designed to separate the bottom 10% from the top 90%. (*Id.*).

"Item discrimination provides an indication of whether an item appropriately discriminates or differentiates between those examinees who perform well on the test and those who perform poorly." (*Id.* at 41). The premise for item discrimination analysis is that "[i]f performance on an item is unrelated to performance on the overall exam, then that item is not differentiating among candidates as intended." (*Id.*). There are two common methods for measuring item discrimination: the index of discrimination and the item-total correlation. The index of discrimination is computed by first dividing test-takers into upper and lower groups based on overall test scores, then subtracting the proportion of the lower group who answered the item correctly from the proportion of the upper group who answered the item correctly. This produces a value, D. One study suggests that questions whose D value is less than two should be eliminated. The item-total correlation "represents the correlation between an item and the rest of the test (i.e., the correlation between the item and the total score on the exam calculated excluding that item)." (*Id.*). A low item-total correlation means that an item has little relationship with the overall test score and does not discriminate between those who perform well and those who perform poorly. (*Id.* at 41). One study concludes that items with an item-total correlation below .05 are "very poorly discriminating item[s]" and that items with an item-total correlation greater than .2 "are at least moderately discriminating." (*Id.* at 42).

plaintiffs, and the HPFFA met to review scoring. Initially, based on the item analysis, the consultants recommended giving candidates full credit for seven job-knowledge questions and for fourteen situational-judgment questions, effectively eliminating those questions as a way to differentiate among the candidates. In addition, the City's internal SMEs recommended giving full credit for one job-knowledge question and for six situational-judgment questions. The panel agreed with the SMEs' recommendation. (Docket Entry No. 94, at 2; Docket Entry No. 94-1 at 2–3). The panel also agreed that a candidate's score on the job-knowledge portion and the situational-judgment portion would be weighted equally in calculating the final score. (Docket Entry No. 94, at 2). Based on these decisions, a rank-order results list for the exam was created.[9] Some discrepancies emerged in the statistical calculations and the consultants recommended a credit adjustment for additional questions. On January 10, 2011, the City submitted a revised rank-order list of candidates who passed the exam. (Docket Entry No. 104, at 2).

On January 12, 2011, the City advised this court that there were more than 200 appeals by the promotional candidates. On January 14, the City moved for additional time to finalize the scores and rank-order list. (Docket Entry Nos. 110 & 112). The City sought more time than the 60 days the TLGC allowed to decide whether to sustain the appeals. The HPFFA did not object to the

[9] One disagreement the panel could not initially resolve was the cut-off score for passing. The HPFFA insisted that the cut-off score remain at 70% as required by the TLGC and the CBA. *See* Tex. Loc. Gov't Code § 143.108(a). The consultants and the City argued that the cut-off score should be calculated according to the "Angoff method." Dr. David M. Finch, an industrial psychologist who consulted with the parties' experts on scoring the exam, stated that the Angoff method is a "professionally accepted process in the field of industrial/organizational psychology" to derive cut-off scores. (Docket Entry No. 94-1, at 3). "[T]he Angoff method requires a panel of SMEs to estimate the percentage of minimally qualified individuals who would correctly answer a particular test question." (*Id.*). Based on these estimates, a cut-off score can be computed. Dr. Finch found that the resulting cut-off score was 68.92%. After conducting further analysis, the consultants determined that whether the cut-off score was 70% or 68.92% made no practical difference in determining who passed the exam and the parties agreed to use a 70% cut-off score. (Docket Entry No. 101, at 2). Dr. Finch concluded that there was no disparate impact from the 70% cut-off score.

request, and this court granted the motion. (Docket Entry No. 117). This court has not been updated on the status of the appeals or on promotions to captain under the November 2010 exam.

2. The May 2011 Senior-Captain Exam and Future Exams

The HPFFA filed its objections to the proposed changes to the May 2011 senior-captain exam and to subsequent captain and senior-captain exams. (Docket Entry No. 89). The proposed changes are described as follows:

1. Officer Development Program
   - 2 years in grade
   - Educational Courses
     - Officer Development I, II
     - Available online at stations for all to participate

2. Job-Knowledge Written Test
   - New job analysis
   - Designed by an industrial/organizational psychologist
   - HFD departmental based
   - Pass/Fail Exam (test designer to determine cut off score)
   - No rank-order list
   - Test designer may elect not to use a written job-knowledge cognitive test

3. Scenario-Based Computer-Objective Test
   - Situational-judgment exam with HFD departmental scenarios
   - Computer simulations, such as in-basket exercises or incident-scenario judgment
   - Zero, partial, and full credit answers may be used
   - Same responses receive same points
   - Scored test
   - Rank-order list from which all of intended promotions will proceed to assessment center

4. Assessment Center
   - Rank-order list
   - Sliding bands based on test accuracy as determined by consultant
   - Fire Chief will document reasons for selection of each candidate within bands
   - No Rule of Three
   - Two-year eligibility list
   - Will be applied to the May 2011 Senior-Captain exam

4a. [Blank]
   - Additional points in effort to minimize disparate impact potential

- All points stay the same as the Collective Bargaining Agreement until May 1, 2011
- The City will propose and will bargain for a point system which does not exceed the following points:
  - 10 points for seniority
  - 5 points for time in rank
  - Education/Certification
    - 1 point-Intermediate certification
    - 2 points-Advanced certification
    - 3 points-Masters certification or Associates degree
    - 4 points-Bachelor's degree
    - 5 points-Masters degree

(Docket Entry No. 69-2, at 29–30).

The parties agree that many of these proposed changes violate the TLGC and the CBA. Under the consent decree, the test designer "may" elect to use a "written job-knowledge test" depending on the job analyses for the captain and senior-captain positions. Whether such questions are included depends on the importance of the "knowledge" component compared to the skills, abilities, and other characteristics identified for the positions. If the test designer elects to use written job-knowledge questions, they are scored on a pass/fail basis. Only candidates who get a passing score remain promotion-eligible. A candidate's specific score on the job-knowledge test is otherwise irrelevant. The score is not used to produce a rank-order list and the Rule of Three is abandoned as to this part of the promotional process.

The remaining two parts of the captain and senior-captain exam are not questions based exclusively on facts and information.[10] One part is a scenario-based computer objective test. The second part uses an assessment center.

[10] *See* TEX. LOC. GOV'T CODE § 143.032(d) (requiring that the exam "test the knowledge of the eligible promotional candidates about information and facts").

The scenario-based computer objective test is based on situational-judgment concepts, using a computer to present hypothetical situations that captains and senior captains would likely encounter on the job. One type of situational-judgment question identified in the consent decree is an "in-basket exercise." In such an exercise, a candidate is given documents or other information creating a hypothetical fact pattern and is asked to analyze or describe a response. An in-basket exercise testing training abilities might ask the candidate to review a firefighter's performance evaluations and identify what training that firefighter needs to improve. (Dr. Brink Report 48). The consent decree allows for scoring these situational-judgment questions on a full-credit, partial-credit, and zero-credit basis, provided that the "same responses" receive the same credit. The consent decree requires ranking the candidates according to their scores on this part. In the initial settlement agreement, the candidates' scores on this situational-judgment component determined whether the candidate could proceed to the final phase of the exam, but the modified settlement agreement provides that all candidates advance. (*Compare* Docket Entry No. 69-2, at 29, *with* Docket Entry No. 86-1, at 3).

The final exam component is an assessment center. "An *assessment center* consists of multiple exercises simulating job activities that are designed to allow trained observers, or assessors, to make judgments about candidates' behaviors as related to job performance." (Dr. Brink Report 47). One type of simulation used in assessment centers is a "role play." A role play "is a simulation of a face-to-face meeting between the candidate (playing the role of a job incumbent) and a trained role player acting as a person incumbents frequently encounter on the job (such as a subordinate or citizen)." (*Id.*). Assessment-center activities such as role play violate the CBA and TLGC. *See* TEX. LOC. GOV'T CODE § 143.032(c) (forbidding tests that "in any part consist of an oral interview"). "Assessors" score promotional candidates' performance on the assessment-center

activities. Although there are preset criteria distinguishing better from worse performance, the scoring system is subjective and violates the TLGC and the CBA.

Another inconsistency between the TLGC and the CBA on the one hand and the consent decree provisions on the other is the requirement in the consent decree to "band" the promotional candidates' assessment center scores. "Banding" scores means adjusting the individual test scores based on statistical analyses showing the likelihood that: (1) a candidate could score higher or lower on the same exam; and (2) the individual assessor could have given the candidate a higher or lower score for the same performance. Banding tends to convert individualized score differences into homogenized "bands" of more uniform scores. For example, three candidates' scores of 85, 86, and 87 might be "banded" as one score of 86, depending on the results of the statistical analysis. Banding is like converting individual scores of 95%, 97%, and 100% on a 100-question multiple choice test into three "As" that are viewed as identical. The conversion is based on statistical analysis showing that an individual scoring 95% on the exam has the same chance of scoring 100% on the exam as the person scoring 100% on the exam has of scoring 95%.[11] Banding is based on the assumption that small differences in scores do not reliably demonstrate superiority in the KSAOs the exam is supposed to measure. One of the City's expert witnesses, Dr. Morris, summarized banding as follows:

> "[A] band is . . . saying if I made 87 and someone else made 85, is it possible that the next day I could have made 85 and they could have made 87? So a band—the band we're trying to calculate the standard error of measurement is simply saying that certain number of times that band is going to fall within a standard error of measurement that we calculate. So, it's a reasonable thing. And most people in our

[11] Dr. Arthur explained that "[a] really good example would be most academic settings, where an A is a hundred to a 90, that's actually a band, but that's an administrative band, not a psychometric band." (Evidentiary Hr'g Tr. 374, Docket Entry No. 131).

> field accept using bands as a way to minimize the error that could be assumed in the minds of the decision-makers.

(Evidentiary Hr'g Tr. 100, Docket Entry No. 130).

Banding is inconsistent with the Rule of Three. Under the proposed consent decree, the final promotion decision is based on the banded assessment-center scores. Names are submitted by score "bands," not subject to the Rule of Three that would have applied under the TLGC and the CBA. Under the Rule of Three, if the three highest scores were 85, 86, and 87, the names of those applicants would be submitted. The person who scored the 87 would be selected unless the decision-maker provided a written reason for selecting the person who scored the 86 or 85. Under the banding system, the three individuals would be treated by the decision-maker as having the same score. The "band" might also be larger than three persons; its size would be determined by statistical analyses rather than a preset number. The consent decree requires the decision-maker to select one within the band and to provide a written explanation for the selection.

The consent decree does not state the role of a candidate's race or how the decision-maker may consider race in choosing who to promote within a band. There was testimony that using race as a factor to select a candidate within a band could reduce the exam's disparate impact on African-American applicants. Within a band, all applicants are viewed as equal. (Evidentiary Hr'g Tr. 107-08, Docket Entry No. 130). But the consent decree does not explicitly authorize race-based promotional decisions.

II. The Evidence in the Record

At an evidentiary hearing, the parties presented evidence as to (1) whether the senior-captain exam disparately impacted black firefighters, and (2) whether the proposed changes to the captain and senior-captain exams were justified by business necessity.

A. The Evidence as to Disparate Impact of Past Exams

On February 8, 2006, the City of Houston administered the senior-captain exam to 221 promotional candidates. Of the 221 candidates taking the exam, 172 were white, 15 were black, 33 were Hispanic, and 1 was "other." The 212 candidates who passed by scoring above 70 consisted of 166 Caucasians, 13 African-Americans, 32 Hispanics, and 1 "other." The City promoted 70 candidates based on the rank-order list of those who passed the exam. Of those promoted, 59 were Caucasian, 2 were African-American, 8 were Hispanic, and 1 was in the "other" category. (Evidentiary Hr'g Ex. 7, Dr. McPhail Report, at 5).

The following experts submitted reports or testified as to whether the 2006 senior-captain exam disparately impacted black firefighters:

- Dr. S. Morton McPhail, an industrial-psychology consultant, on behalf of the City. Dr. McPhail is licensed by the Texas State Board of Examiners of Psychologists and is a Fellow of the Society for Industrial and Organizational Psychology ("SIOP"). He has served as an adjunct faculty member in the psychology departments of Rice University and the University of Houston. Dr. McPhail has "authored scholarly articles and [has] given many symposia, presentations, and continuing education workshops for peers on issues relating to employment, and in several instances, on the topics of job analysis." (Docket Entry No. 37-1, at 3).
- Dr. Kyle Brink, an industrial-psychology consultant and a tenure-track assistant professor in the management department of the Bittner School of Business at St. John Fisher College, testified on the plaintiffs' behalf. Dr. Brink's doctorate is in industrial psychology. He has experience developing and validating promotion procedures at both private companies and governmental organizations. Recently, he

worked with the Personnel Board of Jefferson County, Alabama, helping end a federally imposed consent decree. (Dr. Brink Report 5).

- Dr. Kathleen Lundquist, an industrial-psychology consultant, testified for the City. Dr. Lundquist is the president and CEO of APT, Inc. She has "extensively researched, designed and conducted statistical analyses and provided consultation in the areas of job analysis, test validation, performance appraisal and research design" for "major corporations in the banking, financial services, retail, electronics, aerospace, pharmaceutical, telecommunications, and electric utility industries, as well as for federal, state, and local agencies." (Dr. Lundquist Aff. 1, Docket Entry No. 93-2). She has a Ph.D. in psychometrics from Fordham University.
- Dr. Winfred Arthur, a full professor of psychology and management at Texas A&M University, testified for the HPFFA. Dr. Arthur has a Ph.D. in industrial/organizational psychology from the University of Akron and "over 20 years of practical experience in the areas of test development, selection, public safety testing, and training." He is a SIOP fellow. (Dr. Arthur Aff. 1, Docket Entry No. 89-1).
- Dr. David M. Morris, an industrial-psychology consultant, testified for the City. Dr. Morris is the president of Morris & McDaniel, Inc., an industrial-psychology consulting firm he started in 1976. He received his Ph.D. in psychology, with a specialization in industrial/organizational psychology, from the University of Southern Mississippi. Dr. Morris has authored numerous scholarly articles and is a member of the industrial/organizational division of the American Psychological Association and also a member of SIOP. (Evidentiary Hr'g Ex. 14).

All the experts agreed that the "total selection process" for promoting HFD captains to senior captain showed disparate impact under the 4/5 Rule. The Guidelines require that "[a]dverse impact is determined first for the overall selection process for each job." *Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures*, 44 Fed. Reg. 11966, 11998 (1979) [hereinafter *Guidelines Questions & Answers*]. "The 'total selection process' refers to the combined effect of all selection procedures leading to the final employment decision such as hiring or promoting." *Id.* The experts agreed that the rate of blacks promoted to senior captain—13.3%—is less than 4/5ths the selection rate for whites—34.3%. (*See, e.g.*, Dr. Brink Report 9–10).[12] But only Dr. Brink and Dr. Lundquist found disparate impact for the senior-captain exam.

Both parties' experts testified that a 4/5 Rule violation is an unreliable basis to find disparate impact when the population size of one group is small. Only 17 black firefighters were eligible for promotion to senior captain. The experts agreed that this is too small a number to make a 4/5 Rule violation sufficient to find disparate impact. (Dr. Brink Report 10; Dr. Lundquist Aff. 6; Dr. Arthur Aff. 2–3; Dr. McPhail Report 6–7; Evidentiary Hr'g Ex. 12, Dr. Morris Report, at 0010447). There

---

[12] The Guidelines provide a four-step process for calculating whether there is disparate impact under the 4/5 Rule:

> (1) calculate the rate of selection for each group (divide the number of persons selected from a group by the number of applicants from that group).
>
> (2) observe which group has the highest selection rate.
>
> (3) calculate the impact ratios, by comparing the selection rate for each group with that of the highest group (divide the selection rate for a group by the selection rate for the highest group).
>
> (4) observe whether the selection rate for any group is substantially less (i.e., usually less then 4/5ths or 80%) than the selection rate for the highest group. If it is, adverse impact is indicated in most circumstances.

Guidelines Questions & Answers, 44 Fed. Reg. at 11998.

was general agreement among the experts that when group populations are small, statistical analyses should be used to determine whether the 4/5 Rule violation is the product of "chance." This requires determining the statistical significance of the 4/5 Rule violation. Dr. Morris's report noted that the 4/5 Rule risks "sampling error," which statistical-significance analysis mitigates. Dr. Morris's report stated:

> [T]he 4/5ths Rule has two major limitations, precision and sampling error. The 4/5ths Rule provides a descriptive ratio; it is not a statistical test. As such, the 4/5ths Rule cannot determine if an observed disparity is the result of mere chance or an indication of underlying bias. Use of the 4/5ths Rule is limited further by sample error. Unlike statistical tests, the 4/5ths Rule does not make adjustments for sampling error and, in cases where sample sizes are small, may fail to detect disparities. More problematic, the 4/5ths Rule has proven to falsely show adverse impact when no adverse impact exists.
>
> . . . .
>
> When sample sizes are small, results from the 4/5ths Rule will vary, often dramatically, because the composition of candidates vary each time a personnel decision is made. Statistically speaking, this variation is referred to as sampling error. The 4/5ths Rule is insensitive to sampling error. Boardman (1979) and Greenberg (1979) demonstrated that the 4/5ths Rule is susceptible to either falsely identifying adverse impact when none exists or failing to identify adverse impact when it does exist. At their very heart, statistical tests directly address sampling error.

(Dr. Morris Report 0010446–47). Dr. Brink's report discussed how the 4/5 Rule risks "Type I" error by leading to the conclusion "that adverse impact exists, when in reality the difference in selection rates is a result of sampling error (or chance)." (Dr. Brink Report 12). Dr. Brink's report explained how statistical tests can "control the potential amount of Type I error":

> Type I error in this context is defined as concluding that adverse impact exists, when in reality the difference in selection rates is a result of sampling error (or chance). A statistically significant result is one in which the probability of incorrectly concluding that adverse

> impact exists (i.e., a Type I error) is less than a specified level; this specified level is referred to as an alpha level . . . . Statistical tests produce a probability value . . . that determines or estimates the probability of obtaining the sample result assuming there were no differences in the population. If the [probability value] resulting from the statistical test is less than the specified alpha level, we say the result is statistically significant and would decide, based on the test, that there is adverse impact. For example, if an alpha level of .05 is chosen and the [probability value] resulting from the statistical test is less than .05, then there is less than a 5% probability that the difference is due to chance (i.e., there is less than a 5% probability of making a Type I error) and we say the result is statistically significant. Conversely, you can conclude that there is a 95% probability that the difference is not due to chance.

(*Id.*).

The experts also identified peer-reviewed journal articles critical of the 4/5 Rule. In 1979, Anthony Boardman and Irwin Greenberg authored analyses showing that the 4/5 Rule could lead to both Type I (falsely identifying disparate impact when none exists) and Type II (failing to identify disparate impact when it does exist) statistical errors. *See* Irwin Greenberg, *An Analysis of the EEOCC 'Four-Fifths' Rule*, 25 MGMT. SCI. 762 (1979); Anthony E. Boardman, *Another Analysis of the EEOCC 'Four-Fifths' Rule*, 25 MGMT. SCI. 770 (1979). A recent article similarly concluded that "there is a fairly high false-positive rate for the 4/5ths rule used by itself." Phillip L. Roth *et al.*, *Modeling the Behavior of the 4/5ths Rule for Determining Adverse Impact: Reasons for Caution*, 91 J. APPLIED PSYCHOL. 507, 519 (2006). The article's authors cautioned that "other factors (e.g., sample size) were quite important" and recommended using "a test such as Fisher's exact test or a chi-square test to mitigate false-positives." *Id.* Also noting the 4/5 Rule's shortcomings, Scott Morris and Russell Lobsenz recently proposed a "more complex" statistical technique, the Z_{ir} test,

for evaluating disparate impact. *See* Scott B. Morris & Russell Lobsenz, *Significance Tests and Confidence Intervals for the Adverse Impact Ratio*, 53 PERSONNEL PSYCHOL. 89 (2000).[13]

The experts applied three tests to measure statistical significance: the Fisher exact; the Pearson chi-square; and the Z_d. The Fisher exact test provides "the exact probability of obtaining the observed frequency table (or one more extreme) under the null hypothesis [that no disparate impact exists]" and is particularly suited to analyses involving small sample sizes. Michael W. Collins & Scott B. Morris, *Testing for Adverse Impact When Sample Size Is Small*, 93 J. APPLIED PSYCHOL. 463, 464 (2008). Dr. Brink reported the Fisher exact test probability to be .15—that is, a 15% chance that the observed results were due to pure chance. (Dr. Brink Report 15). Social scientists generally require the p-value—the probability that the observed results are due to pure chance—to be less than .05 for results to be considered statistically significant.

The Pearson chi-square test estimates the probability of obtaining the observed frequency table under the null hypothesis that no disparate impact exists. (*Id.* at 13). Dr. Brink reported the Pearson chi-square probability to be .10—that is, a 10% chance that the observed results were due to pure chance. (*Id.* at 15). Like the Fisher exact test's p-value, the Pearson chi-square p-value is greater than .05 and is viewed as statistically insignificant.

Unlike the Pearson chi-square and Fisher exact tests, the Z_d test does not provide a p-score, the probability that the observed results were due to pure chance. Instead, the Z_d test yields a Z statistic. The difference between the selection rates of the two groups being compared is statistically significant if the absolute value of the Z statistic is greater than 1.96. *See* Collins & Morris, *supra*,

[13] An earlier draft of this article is found at Exhibit 46 of the January 2011 evidentiary hearing.

at 464. Using the Z_d test, Dr. Brink reported a statistically insignificant Z statistic of -1.66. (Dr. Brink Report 15).

Neither the Fisher exact, the Pearson chi-square, nor the Z_d test demonstrated that the 4/5 Rule violation for the total selection process for senior captain was statistically significant. (Dr. Brink Report 12; Dr. Arthur Aff. 3; Dr. McPhail Report 8; Dr. Morris Report 0010448). Only one test of statistical significance, the Z_{ir} test, showed that the 4/5 Rule violation was statistically significant. Unlike the Z_d test, which evaluates the observed difference in selection rates, the Z_{ir} test evaluates the difference in selection-rate ratios. Dr. Brink argued for the use of the Z_{ir} test in these circumstances because the test uses the same comparison as the 4/5 Rule and is "slightly more powerful than the Z_d or chi-square tests, especially as the proportion of minorities is smaller." (Dr. Brink Report 13-14). But the use of the Z_{ir} test is not as well supported in the literature as the other tests, especially in the context of very small sample sizes.[14] One peer-reviewed journal article explained that while the Z_{ir} test was "interesting and deserve[s] greater thought," further research was needed on the test's ability to evaluate disparate impact. *Roth et al.*, *supra*, at 520. The test's creators acknowledged that the Fisher exact test will "provide a more accurate evaluation of statistical significance" than the Z_{ir} test when the smallest expected value in the analysis is less than five. Morris & Lobsenz, *supra*, at 97. Using a formula provided by Morris and Lobsenz, Dr. McPhail calculated the smallest expected value in the analysis to be 4.75 and concluded that use of the Z_{ir} test was inappropriate.[15] (Dr. McPhail Report 9).

---

[14] When asked about the Z_{ir} test at the evidentiary hearing, Dr. Lundquist stated that it was "relatively new to [her]" and that "[i]t's not a common statistic that I've seen in other cases." (Evidentiary Hr'g Tr. 274, Docket Entry No. 130).

[15] Dr. Brink also acknowledged during cross-examination at the evidentiary hearing that a recent article by Morris and Lobsenz advocating the use of the Z_d test did not mention the Z_{ir} test. (Evidentiary Hr'g Tr. 194–95, Docket Entry No. 130). The article was, however, critical of the Fisher exact test. (*Id.* at 197).

Dr. Brink offered other grounds to support his conclusion that under the 4/5 Rule, there was valid evidence of disparate impact. One was the "N of 1" or "flip-flop rule." Dr. Brink stated that the "N of 1 rule calculates an adjusted impact ratio assuming one more person from the minority group . . . and one less person from the majority group were hired (and, consequently, one less minority and one more majority were not hired). If the resulting selection rates are such that the minority selection rate is now larger than the majority selection rate, selection rate differences may be attributed to small sample sizes." (Dr. Brink Report 10).

The Guidelines illustrate the application of the flip-flop rule. The Guidelines present a hypothetical in which 80 Caucasians apply for a job and 20 African-Americans apply for the job. In the hypothetical, 16 Caucasians are hired and 3 African-Americans are hired. The Guidelines use the flip-flop rule to answer the following question: "[i]s evidence of adverse impact sufficient to warrant a validity study or an enforcement action where the numbers involved are so small that it is more likely than not that the difference could have occurred by chance?"

> No. If the numbers of persons and the difference in selection rates are so small that it is likely that the difference could have occurred by chance, the Federal agencies will not assume the existence of disparate impact, in the absence of other evidence. In this example, the difference in selection rates is too small, given the small number of black applicants, to constitute adverse impact in the absence of other information (see Section 4D). If only one more black had been hired instead of a white the selection rate for blacks (20%) would be higher than that for whites (18.7%). Generally, it is inappropriate to require validity evidence or to take enforcement action where the number of persons and the difference in selection rates are so small that the selection of one different person for one job would shift the result from adverse impact against one group to a situation in which that group has a higher selection rate than the other group.

Guidelines Questions & Answers, 44 Fed. Reg. at 11999.

Dr. Brink's report showed that applying the flip-flop rule discussed in the Guidelines to the February 2006 senior-captain exam, the ratio of Caucasians compared to African-Americans selected for promotion did not change. Dr. Brink concluded that this was further evidence that the sample size was not too small to invalidate the disparate-impact evidence provided by the 4/5 Rule.

Dr. Brink also cited the "one-person rule" as additional support for this conclusion. The "one-person rule is computed by taking the difference between actual minority hires . . . and the expected minority hires . . . . If the difference is less than 1, then violations of the 4/5ths rule are likely due to small sample sizes." (Dr. Brink Report 10). When the one-person rule is applied to the results of the 2006 senior-captain exam, the difference between actual minority hires and expected minority hires is not fewer than one. Dr. Brink concluded that "[i]n all cases, the one-person rule indicates that the violations are not due to small samples." (*Id.*).

Dr. Brink's report deemphasized the value of statistical analysis to determine disparate impact for the total selection process for senior-captain promotions. (*Id.* at 15-16). Dr. Brink cautioned against "dogmatic adherence" to the social scientists' use of .05 as the statistical-significance level, stating that the .05 level should not be used in all contexts and noting that lower statistical results can be meaningful. (*Id.*). Statistical analyses estimate only the likelihood that if a different pool of applicants applied for promotion to senior captain, adverse impact would result. But in the context of disparate-impact analysis, the applicant pool is fixed; there is no other relevant pool. Dr. Brink quoted from an article by Collins and Morris stating that:

> when evaluating a promotion decision, the pool of candidates is relatively fixed. If the decision were repeated at a different point in time, the set of candidates under consideration would be mostly the same. In such cases, probabilities based on randomly sampling from a population . . . would not apply. Similarly, probabilities based on random reassignment of participants . . . would not be appropriate.

> Without some theoretical process for producing different patterns of data . . . statistical significance cannot be defined.

(*Id.* at 12). Dr. Brink also testified that research from Collins and Morris suggested that the Fisher exact test was "overly conservative" and recommended abandoning the test as a measure of disparate impact. (Evidentiary Hr'g Tr. 157, Docket Entry No. 130).

Dr. Lundquist analyzed data relating to the senior-captain promotion process beyond the 2006 senior-captain exam.[16] The data is summarized in the table below. The proportion of white candidates promoted to black candidates is the Adverse Impact Ration ("AIR") set out in the last column. The 4/5 Rule requires an AIR less than .8 to show disparate impact.

	Black Candidates	**White Candidates**	**Black Candidates Promoted**	**White Candidates Promoted**	**AIR**
1993	17	147	0	36	0.00
1996	13	84	1	21	0.31
1999	12	122	0	31	0.00
2002	8	104	2	64	0.41
2006	15	172	2	59	0.39
2009[17]	14	136	2	31	0.63
Overall	79	765	7	242	

Dr. Lundquist testified that a historical analysis using data from multiple test administrations provides "a much more accurate picture of adverse impact of a promotional process." (Dr. Lundquist Aff. 6, Docket Entry No. 93-2). Dr. Arthur, however, was dismissive of any historical

[16] Dr. Lundquist acknowledged that the exams for these promotional cycles asked different questions, but her understanding was that "what you had across those six occasions were essentially six parallel forms of the same kind of multiple-choice questions based on books that people took." (Evidentiary Hr'g Tr. 279, Docket Entry No. 130).

[17] Dr. Lundquist did not have the complete data for promotions to senior captain when she performed her calculations.

aggregation approach because the data would necessarily extend beyond the 2006 exam that was at issue. Dr. Arthur did state that if one were to look to historical data, "the correct analysis would be a Mantel-Haenszel chi-square test." (Dr. Arthur Aff. 3, Docket Entry No. 89-1). Dr. Lundquist agreed with Dr. Arthur that the Mantel-Haenszel test was the most methodologically appropriate analysis. The Mantel-Haenszel test allows statisticians to investigate the consistency of data trends over time while avoiding errors due to aggregation. Applied to the senior-captain exams, Dr. Lundquist found that the Mantel-Haenszel test showed a statistically significant pattern of adverse impact against African-Americans.

B. The Validity Evidence

1. The City's Job Descriptions

The City has not conducted a full validity study of its promotional exams. (Dr. Brink Report 20). The City has, however, produced job descriptions for both the captain and senior-captain positions. (Docket Entry No. 96-1, at 54). The job descriptions contain a detailed listing of the responsibilities for each. The responsibilities for the captain position include: (1) supervising "the emergency response of their assigned apparatus to ensure a safe and timely response to alarms"; (2) supervising "the proper development of an adequate water supply when ordered to do so or fire conditions dictate"; (3) providing "leadership in performing search and rescue operations"; (4) presenting "programs to the community on safety and fire prevention topics"; (5) assisting "in training new employees"; (6) maintaining "proper staffing for all apparatus assigned to their fire station"; and (7) maintaining "the personnel records of members assigned to their station and shift."

Under the "specifications" section are subheadings for "basic knowledge," "specific knowledge," "advanced skills," and "ability to." Bullet points beneath the subheadings describe the KSAOs for captains. The "basic" and "specific" knowledge identified for captains include: (1)

knowledge directly related to firefighting—such as knowledge about municipal and private fire protection, building construction, and water supplies; (2) knowledge about federal, state, and local laws; (3) knowledge about HFD operational procedures; and (4) other types of knowledge, such as about Robert's Rule of Order. (Docket Entry No. 96-1, at 56). The advanced skills include management and supervision, organization, problem-solving, firefighting strategy and tactics, teaching, and public relations. (*Id.*). The abilities include: communication abilities, such as establishing and maintaining working relationships with subordinates as well as "impromptu public speaking"; tactical abilities, such as implementing firefighting strategies; leadership abilities, such as recognizing and responding to individual and group needs; and administrative abilities. (*Id.* at 57). The senior-captain KSAOs are similarly organized and described.

2. Dr. James C. Sharf's Report

Dr. James C. Sharf prepared a report for the HPFFA concluding that the promotional exams are valid based on "validity generalization," a validation method not described in the Guidelines. Dr. Sharf, an employment consultant specializing in risk management, has published "a dozen professional publications including two peer-reviewed chapters: 1) in The Society for Industrial and Organizational Psychology Practice Series (2000), and 2) in American Psychological Association Books (2008)." Dr. Sharf served as Special Assistant to the Chairman of the EEOC from 1990 to 1993 and as the EEOC's chief psychologist from 1974 to 1978. Dr. Sharf has "over three decades' experience developing, implementing and defending selection and appraisal systems in both the public and private sector." He is a Fellow of the Society for Industrial and Organizational Psychology, a Fellow of the Association for Psychological Science, and a Fellow of the American Psychological Association. (Dr. Sharf Report 2, 5).

Acknowledging that the Guidelines identify content-related, criterion-related, and construct-validity studies as proper validation methods, Dr. Sharf argued that these studies establish only a starting point for validity analysis. Dr. Sharf pointed out that the EEOC did not intend the Guidelines to preclude "other professionally acceptable techniques with respect to validation of selection procedures" because test-validity science has evolved since the Guidelines' publication. 29 C.F.R. § 1607.14; *see also Guidelines Questions & Answers*, 44 Fed. Reg. at 12002 ("The validation provisions of the Guidelines are designed to be consistent with the generally accepted standards of the psychological profession."). Dr. Sharf also pointed out that since the Guidelines were published, the American Psychological Association has revised the Standards for Educational and Psychological Tests ("APA Standards") and the Society for Industrial and Organizational Psychology has revised the Principles for the Validation and Use of Personnel Selection Procedures ("SIOP Standards"). (Dr. Sharf Report 8). In light of the revisions to the APA and SIOP Standards, and recent trends in industrial-psychology scholarly publications, Dr. Sharf's report concludes that validity generalization shows that the captain and senior-captain exams are valid, and that validity generalization better analyzes the validity of a test than the methods identified in the Guidelines.

Dr. Sharf argues that one basis for deemphasizing the Guidelines' validation methods is the criticism by industrial psychologists of the Guidelines' recommended approach to job analysis. Dr. Sharf characterized the Guidelines' job analysis as requiring a detailed list of job tasks based on "observable behaviors." Dr. Sharf's report identified several scholarly articles published in psychology journals suggesting that job analyses based on detailed task descriptions are unreliable or unhelpful. For example, a 1981 article in the *Journal of Applied Psychology* by Schmidt, Hunter, and Pearlman, found that detailed job analyses based on observable behaviors created the appearance of large differences between jobs "that are not of practical significance in selection."

(*Id.* at 11). Similarly, the SIOP Standards require only a "general" description of KSAOs and allow a "less detailed analysis . . . when there is already information descriptive of the work." (*Id.* at 15).

A second basis for Dr. Sharf's criticisms of the Guidelines was their emphasis—reflected, for example, in the job-analysis provisions—on "observable behaviors." Dr. Sharf contrasted "observable behaviors" with "unobservable cognitive skills," which the Guidelines do not emphasize. Dr. Sharf argued that the Guidelines' focus on validity studies measuring observable behaviors produces less reliable results than validity studies measuring cognitive skills. Dr. Sharf's report cited a number of scholarly articles finding that a promotional candidate's cognitive skills better predict performance after promotion than observable behaviors. (*Id.* at 30–31). He summarized the articles' findings as follows: "The conclusion from these studies is that pencil and paper tests of cognitive ability such as verbal, quantitative and technical / problem solving abilities not only predict job performance but that they predict job performance better than any alternative—the general case of validity generalization research empirically built upon measures of cognitive ability." (*Id.* at 32). Dr. Sharf acknowledged that some job learning occurs for any candidate, but he argued that recent research shows that individuals with greater cognitive ability will acquire the skills necessary to perform a job successfully. (*Id.* at 36–37). In light of these studies, Dr. Sharf concluded that the Guidelines' "emphasis on 'observable behavior' is both illogical and out of touch with contemporary industrial psychology because there is no knowledge, skill or ability which does not depend on unobservable mental processes involving cognitive abilities." (*Id.* at 11).

Dr. Sharf urged that the more reliable method for analyzing validity is validity generalization. "Validity generalization is industrial psychology's science of the general case demonstrating empirically that the cognitive abilities most studied in industrial psychology—verbal,

quantitative and technical abilities—are also the best predictors of job performance." (*Id.* at 14). Generally, validity generalization analyzes whether a test reliably measures the verbal, quantitative, and technical skills a job requires rather than its KSAOs. The SIOP Standards recognize validity generalization as one method for validating a cognitive-based test. A former EEOC senior attorney has also argued that validity generalization is a valid measure of a test's validity. (*Id.* at 13).

Dr. Sharf used validity generalization to analyze the captain and senior-captain exams. Dr. Sharf first analyzed whether the HFD's job descriptions for the captain and senior-captain positions were valid as a job analysis for validity generalization. His focus was on whether the job descriptions reflected the "verbal, quantitative and technical/problem solving cognitive abilities" required for the positions. (*Id.* at 16). Describing the HFD job descriptions as "among the most comprehensive, thorough, succinctly described job descriptions that I have ever come across," (*id.*), Dr. Sharf categorized the responsibilities and the KSAOs identified in the job descriptions into verbal, quantitative or technical/problem-solving skills. For example, Dr. Sharf classified a senior captain's responsibility to assist "in department supervisory and administrative activities as assigned" as a "verbal ability," (*id.* at 17); a senior captain's responsibility to maintain "inventories of all station, apparatus, and equipment" as a "quantitative ability," (*id.* at 19); and a senior captain's responsibility to assume "command at Emergency Operations incidents" as a "technical/problem solving ability," (*id.* at 20).

Dr. Sharf noted that the additional responsibilities identified in the HFD job description for the senior-captain position related to assuming control, maintaining control, coordinating, supervising, and evaluating the "most effective use." (*Id.* at 17). A captain is responsible for "siz[ing]-up the scene at emergency medical calls, as a first responder, in order to begin providing needed emergency medical intervention to mitigate the problems encountered within the HFD

guidelines." (*Id.* at 19). A senior captain is also responsible for assuming "control of medical emergencies when arriving first," maintaining "control of patient care until the arrival of a higher medical authority," and supervising or performing "medical intervention in accordance with one's level of training." (*Id.* at 19). Based on the emphasis on control and supervision in the HFD job description and a 1986 article published by Hunter & Hunter in *Psychological Bulletin* suggesting that "the more complex the job, the better cognitive ability predicted job performance," Dr. Sharf concluded that the senior-captain position requires greater cognitive ability than the captain position.

Dr. Sharf also compared the City's job descriptions to the job analysis for municipal firefighters created by the United States Department of Labor (the "DOL Analysis").[18] Dr. Sharf found that the HFD's captain and senior-captain job descriptions were consistent with the DOL Analysis. Based on these similarities and "generally accepted principles and practices of industrial psychology," Dr. Sharf concluded that the HFD had conducted a valid job analysis for developing an "objective, job-related Captain's and Senior Captain's exam." (*Id.* at 22).

Using the recategorized HFD job descriptions and the DOL Analysis, Dr. Sharf prepared a "combined job analysis." (*Id.* at 22–30). Dr. Sharf then discussed whether the HFD captain and senior-captain exams validly measured the cognitive skills identified in the combined job analysis and found that they did. Dr. Sharf's report does not provide the analysis that led to his conclusion or refer to statistical studies to provide support. Dr. Sharf relies on scholarly articles arguing that individual differences in cognitive performance correlate to job performance to support his conclusion that the captain and senior-captain exams are valid under validity generalization.

[18] The DOL Analysis is available online through O*NET. Like the HFD analysis, the DOL Analysis is organized by KSAOs. O*NET indicates that the DOL Analysis was based on ratings by subject-matter experts on a five-point scale of the responsibilities and KSAOs associated with municipal firefighting. (Dr. Sharf Report 21–22).

Dr. Sharf's expert report also criticized noncognitive measures of test validity. He argued that video- or situational-based assessments are poor simulations of actual situations a captain or senior captain encounters. He colorfully explained that "[a] video depiction is hardly the stress of an adrenalin rush from the danger of a whiff of noxious chemicals or a lung full of searing smoke." (*Id.* at 37). Dr. Sharf also emphasized that the knowledge a candidate brings to the position of captain—"what you think"—will impact how he responds to emergency situations. He argued that having the technical knowledge required to respond to certain emergency situations is a prerequisite to making the appropriate response.

3. Dr. McPhail's Report

Dr. McPhail conducted a criterion-related validation of the 2006 captain exam. He did not conduct a similar validation study for the 2006 senior-captain exam. Dr. McPhail's criterion-related validation study compared candidates who were promoted to captain based on the 2006 exam with those who were not but who "rode up" as captain after the exam. The validation study analyzed whether there was a relation between success on the exam and performance as captain by comparing the promotional candidates' 2006 exam scores with performance evaluations created by Dr. McPhail and filled out by supervisors. Dr. McPhail concluded that the validation study showed only "equivocal" results as to the captain exam's validity.

Initially, Dr. McPhail identified "a set of performance dimensions appropriate and important to effective performance as a Captain," using the HFD's job description for the position, the exam, and information from internal SMEs. (Docket Entry No. 37-1, at 6). Dr. McPhail's report identified nine specific performance dimensions: "emergency operations"; "station management"; "technical knowledge"; "management resources"; "supervision"; "problem solving"; "interpersonal

effectiveness"; "professional orientation & commitment"; and "overall job performance."[19] (*Id.* at 37). After more discussions with the SMEs, Dr. McPhail developed "performance evaluation behavioral anchors" related to each dimension. Dr. McPhail then asked the SMEs to evaluate each behavioral anchor's relative importance. Based on these evaluations, the behavioral anchors were classified according to a five-point rating scale. For example, a five-point "exemplary rating" for "emergency operations" requires the following behavioral anchors: applying appropriate triage to prioritize transport of injured persons, quickly sizing up "the scene and us[ing] resources to prioritize evacuation of threatened occupants of multi-story building while simultaneously maintaining communication with IC and initiating interior fire attack," and identifying "immediate rescue situation in burning building and prioritiz[ing] deployment of truck line to protect tactically correct extrication of victim prior to laying supply line." (*Id.* at 38). A three-point "meets job requirements" rating requires only one behavioral anchor: laying "proper supply line before entering building to search for unknown possible victims and ensure safety of fire fighters." (*Id.*).

Using the performance dimensions, the behavioral anchors, and the SME evaluations, Dr. McPhail created a Performance Dimension Rating Form ("PDRF"). (*Id.* at 40). Each PDRF asks a supervisor to analyze one performance dimension of the individual to be scored. The performance dimension is identified at the top of the PDRF. The five rating categories are placed on a left-hand column and the behavioral anchors are listed beneath each. Within each rating category are twelve

---

[19] Dr. McPhail's report further described each performance dimension. For example, the technical knowledge is described as follows:

> Applies knowledge of safe work practices, fire fighting technology, fire science, emergency scene management, HFD Standard Operating Guidelines, and legal requirements to maximize safety of the public and HFD personnel and preservation of property.

(Docket Entry No. 37-1, at 37).

possible scores. The possible scores start at one, which corresponds to the "unacceptable" rating category—the lowest rating possible—and end at sixty, which corresponds to the "exemplary" rating. (*Id.*). An individual whose performance in emergency operations is "unacceptable" can receive a score from one to twelve and an individual whose performance is exemplary can receive a score from forty-nine to sixty. (*Id.*).

The PDRFs were circulated to HFD district chiefs, who supervise captains. The district chiefs were asked to score the performance dimensions for captains promoted after the January 2006 captain exam and EOs who were not selected for captain after the exam but had ridden up as captains after that. In January 2006, 438 firefighters took the exam and 157 were promoted to captain. Of those who took the exam but were not promoted, 281 EOs rode up as captains. Of the 84 district chiefs asked to evaluate performances, 77 submitted evaluations. The results of the supervisors' assessments of the captains and EOs was then compared to the scores on the January 2006 captain exam. (*Id.* at 60). The raw data for the PDRFs showed a mean score of 79.35, with a 10.13 standard deviation, for the 438 firefighters who took the January exam; a mean score of 90.10 with a 3.83 standard deviation, for the 157 firefighters promoted to captain based on the January exam; and a mean score of 73.35, with a 7.13 standard deviation, for EOs who took the January exam but who were not promoted and later rode up as captains. (*Id.* at 45).

Dr. McPhail constructed a "validation sample" of 199 firefighters to test the validity of the 2006 exam. Using the validation sample, Dr. McPhail conducted statistical analyses "to evaluate evidence for the validity of the promotional examination." (*Id.* at 52). The analyses included "zero-order (bivariate) correlations for three different samples: the entire validation sample, only those promoted to captain, and only those not promoted to captain." (*Id.*). Dr. McPhail found that the bivariate correlations "appeared to provide supporting evidence for the validity of the

examination." He noted that the correlations "were all significant and ranged from r = .37 to r = .51." (*Id.*). But when Dr. McPhail placed the results of the bivariate analysis on a scatter plot, he noted that the relationships between the "examination scores and criteria indicated barbell shaped bivariate distributions, in which most of the performance ratings for captains were located in the upper end of the distribution and most of the performance ratings for engineers/operators were located in the lower end of the distribution." (*Id.*). He explained that this supported at least two inferences: (1) "the captain promotional examination effectively taps the intended construct domain which results in the observation that those scoring higher on the exam tend to have higher performance"; and (2) "because promotional examination scores were used as a basis for promotion . . . scores should be correlated with captain performance because those at the formal captain rank have a greater opportunity to acquire knowledge and skills integral to effective functioning." (*Id.*).

Dr. McPhail also separated the validation sample into those who were promoted to captain and the EOs who were not. He then conducted bivariate correlations between the exam scores and the performance criteria for each group. For the captain sample, there were no significant correlations between the promotional exams and the performance criteria. For the EO sample, the results were "inconsistent." Dr. McPhail found that the promotional exam correlated to station management, management of resources, and problem solving, "but were not significantly correlated with other criteria." (*Id.* at 54).[20]

Dr. McPhail also used "a number of multiple regression models" to measure the January 2006 captain exam's validity. He explained these analyses, as follows:

[20] As a precaution against potential rater effects— rating tendencies that are idiosyncratic to each district chief providing the ratings—"consultants standardized the performance ratings within District Chief." (Docket Entry No. 37-1, at 54).

> For each criterion, a variable identifying individuals as either captains or [EOs] and the captain promotional examination variable were entered as predictors in the regression equation. If the regression weight associated with the promotional examination variable was significant, the results suggested that the promotional examination predicts the respective criterion variable above and beyond the prediction provided by the variable identifying rank. That is, the effect of being a captain versus an [EO] on the criterion variable is accounted for, and any additional prediction provided by the promotional examination variable can be interpreted as the unique impact of promotional examination scores.

(*Id.* at 55). The regression analyses showed that the January 2006 captain exam significantly predicted station management, resource management, and problem solving. (*Id.*).

Dr. McPhail concluded that the statistical analyses showed "equivocal evidence of the predictive capability of the 2006 examination." (*Id.* at 60). He noted that the analyses of the entire sample showed "substantial and statistically significant correlations . . . between the test and rated performance," but that the bivariate scatter plot moderated the correlations. (*Id.*). He also noted that both the bivariate analysis and the multiple-regression analysis showed significant correlations within the EO subgroup with station management, management of resources, and problem solving. Dr. McPhail concluded that "among a much less restricted sample, test scores provided incremental prediction of performance . . . even after accounting for the relationship of promotion status with the performance ratings." (*Id.* at 61).

4. Dr. Brink's Report and Testimony

Dr. Brink's report, prepared for the City, used the Guidelines, SIOP Standards, and scholarly articles to criticize the 2006 captain and senior-captain exams. Dr. Brink's report criticized the job descriptions offered by the HFD as job analyses, the "linkage" between the HFD's job descriptions and the exam, the reliability of the exam, and the processes the City used to establish the promotional system. Based on these criticisms, Dr. Brink concluded that the captain and senior-

captain exams were not content-valid and that the City's promotional process violated the Guidelines. Dr. Brink's expert report also identified alternative evaluation measures.

Dr. Brink's report identified numerous shortcomings in the HFD job analyses. The report explained that both the Guidelines and SIOP Standards emphasize the importance of a job analysis to determine content validity. The Guidelines contain detailed requirements for the job analysis prepared for a content-validity study. *See generally* 29 C.F.R. § 1607.15(C). Similarly, the SIOP Standards state that "[e]vidence for validity based on content rests on demonstrating that the selection procedure adequately samples and is linked to the important work behaviors, activities, and/or worker . . . knowledge, skills, abilities, and other characteristics . . . defined by the analysis of work." (Dr. Brink Report 20). The report proceeded to detail the weaknesses of the City's job descriptions as job analyses in light of the Guidelines and the SIOP Standards.

The report first discussed the City's failure to maintain records of "required information" for documenting validity. Dr. Brink described the records the City produced to document the promotional tests' validity as "almost 8,000 mishmash pages." (*Id.* at 20). Under the Guidelines, the City should record the users, locations, and dates of a job analysis and any purposes related to the analysis. 29 C.F.R. § 1607.15(C)(1), (2).[21] The City did produce questionnaires used to develop the job descriptions, the job descriptions themselves, and "incumbent frequency ratings, supervisor criticality ratings, and computer overall criticalities" of the KSAOs identified in the job descriptions. (Dr. Brink Report 23). Dr. Brink, however, found that these documents did not adequately validate the City's promotional processes for the captain and senior-captain positions.

[21] Dr. Brink's report also stated that the City failed to document the bases for including the source materials selected for the 2006 exams. (Dr. Brink Report 29).

Dr. Brink's report concluded that the job descriptions based on questionnaire responses were insufficient as job analyses. A job analysis should incorporate information from a number of sources, including background research, observing SMEs performing the job, interviewing SMEs, SME focus groups, and job-analysis questionnaires. The more sources are incorporated, the stronger the job analysis will be. (*Id.* at 21). Dr. Brink found that the City relied exclusively on questionnaires. He noted that the job descriptions and questionnaires the City supplied stated that "any one position may not include all the duties listed, nor do the examples listed necessarily include all duties performed." (*Id.*). Dr. Brink also noted that it was not clear whether internal SMEs had participated in creating the job descriptions. Dr. Brink argued that the lack of SME input would raise concerns about accuracy. Dr. Brink noted as an example that a specific knowledge identified in the captain job description was "Robert's Rule of Order," but that "several" captains did not know what this meant. (*Id.* at 22). Finally, Dr. Brink's report emphasized the vagueness of both the questionnaires and the job descriptions. For example, the job descriptions for both captain and senior captain list "training" as a "knowledge." (*Id.*). Dr. Brink argued that these descriptions failed to meet the Guidelines' requirement that "an operational" definition should be provided for each KSAO. 29 C.F.R. § 1607.15(C)(3).

Dr. Brink's report also found that the City's "incumbent frequency ratings, supervisor criticality ratings, and computer overall criticalities" of the KSAOs identified in the job descriptions showing their relative importance did not correlate with the questions asked on the captain and senior-captain exams. His criticisms for the captain job-description assessments included the following:

- 'Responsibilities' AA and BB (the third and fourth most critical responsibilities) were not assessed by any items[;]

> - 'Responsibilities' P and Q (the 2 *least* critical of the 35 responsibilities) were each . . . assessed by 12 items; more items than any of the five most critical responsibilities[; and]
>
> - Sixteen of the 18 'specific knowledges' were not assessed by any items; four of them . . . had some of the highest criticalities.

(Dr. Brink Report 23). He had similar criticisms for the senior-captain job-description assessments:

> - 'Responsibility' N (the fifth lowest criticality) was . . . assessed by 25 items while AA and BB (the third and fourth most critical responsibilities) were assessed by 13 items each and X (the sixth most critical) was assessed by only 1 item[;]
>
> - The three most critical 'specific knowledge[s]' . . . are . . . assessed by 0, 16, and 2 items . . . whereas other less critical specific knowledge is . . . assessed by as many as 36 items[; and]
>
> - Six of the nine 'advanced skills' . . . were . . . assessed by 39 or 40 items regardless of criticality.

(*Id.*).

After consulting with SMEs, Dr. Brink found that 63% of the captain exam content and 86% of the senior-captain exam content did not reflect knowledge or skills necessary for the first day of work. (*Id.* at 24). Dr. Brink found that this was evidence that the test violated the SIOP Standards requirement that a "selection procedure should be based on an analysis of work that defines the balance between the work behaviors, activities, and/or [KSAOs] the applicant is expected to have before placement on the job." (*Id.* at 24). The Guidelines similarly require that "[f]or any selection procedure measuring a knowledge, skill, or ability the user should show that (a) the selection procedure measures and is a representative sample of that knowledge, skill, or ability; and (b) that knowledge, skill, or ability is used in and is a necessary prerequisite to performance of critical or important work behavior(s)." 29 C.F.R. § 1607.14(C)(4). Referring to this as the "necessary-upon-

promotion" requirement, Dr. Brink concluded that the test poorly assessed whether a promotional candidate has the KSAOs required to begin work as a captain or senior captain.

According to Dr. Brink, "[p]erhaps the most condemning fact regarding the job analysis is that it was completely irrelevant." (Dr. Brink Report 26). The report stated:

> The exams were developed based on the information provided in the text books that were on the 'source materials list.' The source materials lists for Captain and Senior Captain were established on August 17 and September 20, 2005, respectively. The Captain and Senior Captain jobs were announced . . . on October 4 and October 26, 2005, respectively. It is not documented when the Captain job analysis questionnaire was commenced or completed; however, it was still ongoing well into October . . . . Therefore, the job analysis . . . was not even completed for Captain or even started for Senior Captain until after the source materials lists were determined and announced to candidates. . . . This backwards approach to test validation is clearly inappropriate and invalid . . . .

(*Id.*).

Dr. Brink's report also criticized the exam itself, concluding that the "linkage" of test questions to the captain and senior-captain positions was "too abstract." (*Id.* at 31). This was inconsistent with the Guidelines, which state as follows:

> There must be a defined, well recognized body of information, and knowledge of the information must be prerequisite to performance of the required work behaviors. The work behavior(s) to which each knowledge is related should be identified on an item by item basis. The test should fairly sample the information that is actually used by the employee on the job, so that the level of difficulty of the test items should correspond to the level of difficulty of the knowledge as used in the work behavior.

Guidelines Questions & Answers, 44 Fed. Reg. at 12007. Dr. Brink stated that the identified responsibilities should have been linked to KSAOs and that the responsibilities and KSAOs should in turn have been linked to specific exam questions. He noted that there was no documentation linking the source materials to individual questions. Although the City provided "matrices linking

responsibilities and [KSAOs] to source material lists," Dr. Brink found that many of the matrices did not correlate with the cited portions of the source materials. (Dr. Brink Report 31). The City also provided documents linking responsibilities and KSAOs to the exam questions, but Dr. Brink found that the questions rarely correlated with the responsibilities and KSAOs. Dr. Brink gave the following examples:

> For Senior Captain, responsibilities E (search and rescue) and M (assumes command) and basic knowledge K (safety accident prevention) and L (incident management systems) each were supposedly assessed by over half of the exam . . . and all were supposedly assessed by items 1–45 and 71–80. Many of these questions have nothing to do with any of these responsibilities/KSA[O]s (e.g., question 25 asks about the primary cause of cardiac arrest in infants and children), much less assess all of these as well as the many other abilities that are supposedly assessed by them.

(*Id.* at 32).

Dr. Brink also faulted the exam for failing to assess the identified KSAOs in "the context in which they are used on the job." (*Id.*). The basis for this conclusion was Dr. Brink's experience that objective multiple-choice exams poorly evaluate supervisory, leadership, and communication skills and that such exams fail to simulate situations using job-related abilities. During the evidentiary hearing, Dr. Brink distinguished between "high fidelity" tests, which closely resemble actual job behaviors, and "low fidelity" tests, which do not resemble job behaviors. Dr. Brink testified that multiple-choice tests are "low fidelity" and that only high-fidelity tests are likely to have content validity. (Evidentiary Hr'g Tr. 149, Docket Entry No 130).

Dr. Brink's report also stated that low-fidelity tests are inconsistent with the Guidelines. The Guidelines state that:

> The closer the content and the context of the selection procedure are to work samples or work behaviors, the stronger is the basis for

> showing content validity. As the content of the selection procedure less resembles a work behavior, or the setting and manner of the administration of the selection procedure less resemble the work situation, or the result less resembles a work product, the less likely the selection procedure is to be content valid, and the greater the need for other evidence of validity.

29 C.F.R. § 1607.14(C)(4). Similarly, the Q&As in the Guidelines note that:

> Paper-and-pencil tests which are intended to replicate a work behavior are most likely to be appropriate where work behaviors are performed in paper and pencil form (e.g., editing and bookkeeping). Paper-and-pencil tests of effectiveness in interpersonal relations (e.g., sales or supervision), or of physical activities (e.g., automobile repair) or ability to function properly under danger (e.g., firefighters) generally are not close enough approximations of work behaviors to show content validity.

Guidelines Questions & Answers, 44 Fed. Reg. at 12007.

Dr. Brink identified a number of questions on both the captain and senior-captain exams to illustrate his arguments against the exclusive use of multiple-choice questions. (Dr. Brink Report 33). A candidate's ability to delegate is examined by asking for the definition of "delegate"; a candidate's ability to manage a station is examined by asking about the difference between "managers" and "leaders"; and a candidate's ability to ensure the safety of personnel is measured by asking about the definition of "human factors theory of accident causes." (*Id.*). Dr. Brink summarized this aspect of his findings about the exams:

> There is no evidence that memorization of trivial management and diversity facts and definitions is related to success as Captain or Senior Captain. . . . The exams assess the ability to learn a body of information; at best, this is only one of many determining factors with respect to success as a Captain or Senior Captain. The exams also assess characteristics such as reading comprehension, test-taking ability, motivation to study, ability to memorize, leisure time, and disposable income (Captain source material cost $217.29 and Senior Captain source materials cost at least $177.23 . . .). These characteristics are all important for success in school; however, none of these characteristics are included in either job analysis.

(*Id.* at 34).

Dr. Brink also examined the City's use of time limits and a cutoff score, finding no validity support for either. As to time limits, the Guidelines state that "[e]stablishment of time limits, if any, and how these limits are related to the speed with which duties must be performed on the job, should be explained." 29 C.F.R. § 1607.15(C)(5). As to cutoffs, the Guidelines state the following:

> The methods considered for use of the selection procedure (e.g., as a screening device with a cutoff score, for grouping or ranking, or combined with other procedures in a battery) and available evidence of their impact should be described (essential). This description should include the rationale for choosing the method for operational use, and the evidence of the validity and utility of the procedure as it is to be used (essential). The purpose for which the procedure is to be used (e.g., hiring, transfer, promotion) should be described (essential). If the selection procedure is used with a cutoff score, the user should describe the way in which normal expectations of proficiency within the work force were determined and the way in which the cutoff score was determined (essential). In addition, if the selection procedure is to be used for ranking, the user should specify the evidence showing that a higher score on the selection procedure is likely to result in better job performance.

Id. § 1607.15(C)(7). The Guidelines also state that:

> Where cutoff scores are used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force. Where applicants are ranked on the basis of properly validated selection procedures and those applicants scoring below a higher cutoff score than appropriate in light of such expectations have little or no chance of being selected for employment, the higher cutoff score may be appropriate, but the degree of adverse impact should be considered.

Id. § 1607.5(H).

Dr. Brink also found that the City violated the Guidelines and SIOP Standards requirement that an employer consider alternative measures for promotional systems. His report stated that both the Guidelines and the SIOP Standards require an employer, after conducting a job analysis, to

determine how to assess the KSAOs identified. Because there was no evidence that the City considered any format other than an objective, multiple-choice test, Dr. Brink concluded that the City's approach was inconsistent with the SIOP Standards and Guidelines. (Dr. Brink Report 29).

Using the City's data on test scores, Dr. Brink also reviewed the "item analysis" the City conducted. An "item analysis" measures a test's reliability by looking at its measurement error. "[F]or an employment test to accurately predict job performance, it must be reliable; but having a reliable test does not guarantee accurate prediction of job performance." (*Id.* at 39). Dr. Brink measured reliability using a formula proposed in an article by Ghiselli, Campbell, and Zedeck. The formula produces a "validity coefficient" that ranges from 0 to 1. The higher the coefficient, the more valid the test. (*Id.* at 38). Relying on an article by Nunnally and Bernstein, Dr. Brink acknowledged that "the level of reliability that is considered satisfactory depends on how a test is being used," but that in all cases, the validity coefficient should be at least .70. (*Id.* at 39). Dr. Brink stated that for "high stakes testing" like the captain and senior-captain exams, which provide the most important aspect of the promotional decision, the validity coefficient should be .90 at a "bare minimum," and .95 is "desirable." (*Id.*).

The City has conducted "some" item analysis for the captain exam and has conducted an item analysis for 7 items on the senior-captain exam. The City's item analysis showed validity coefficients ranging from .458 to .646 for criteria measured on those exams. Dr. Brink criticized the City's item analysis because it analyzed broad criteria. For example, on the senior-captain exam, the City measured validity against the following criteria: "strategic & tactical considerations on the fireground" (items 1–20); "fire service first responder" (items 21–45); "supervisor" (items 46–70); "terrorism response" (items 71–80); and "Houston Fire Department Guidelines" (items 81–100). (*Id.*). Dr. Brink's report stated that the City's item analysis failed to include any analysis based on

the specific KSAOs within each large criteria group. Dr. Brink also stated that the relevant literature shows that measuring large numbers of items at the same time increases the validity coefficient. The City measured the validity of all 100 test items instead of measuring the test validity within each subcriteria, which would inflate the validity coefficient while failing to capture important criteria.

Dr. Brink also measured the "item difficulty" of the exams. "Item difficulty is a statistic used in test analysis that indicates the percentage of applicants who answered an item correctly." (*Id.* at 40). Item difficulty measures reliability, not validity. Dr. Brink stated that "[t]he purpose of a valid promotional exam is to differentiate candidates based on job-related criteria; if all or most candidates get an exam question correct or incorrect, the item is useless for this purpose." (*Id.*). Items with difficulties above .9 (90% applicants answered correctly) should be eliminated unless the exam is designed to separate the bottom 10% of applicants from the top 90%. Dr. Brink found that 34 items on the captain exam had item difficulties over .9 and that 60 items for the senior-captain exam had item difficulties above .9.

Another way to measure whether an item differentiates between candidates who will perform well after promotion and candidates who will not is through "item discrimination." As the label indicates, item discrimination "provides an indication of whether an item appropriately discriminates or differentiates between those examinees who perform well on the test and those who perform poorly . . . . If performance on an item is unrelated to performance on the overall exam, then that item is not differentiating among candidates as intended." (*Id.* at 41).

The two common methods for measuring item discrimination are the index of discrimination and the item-total correlation. The index of discrimination is computed by first dividing the examinees into upper and lower groups based on overall test scores, then subtracting the proportion of the lower group who answered the item correctly from the proportion of the upper group who

answered the item correctly. This produces a value, "D." Crocker and Algina argue that questions with a "D" value lower than .2 should be eliminated. Dr. Brink found that the captain exam had 53 items with a "D" value below .2 and the senior-captain exam had 66 items with a "D" value below .2. (*Id.*).

The item-total correlation "represents the correlation between an item and the rest of the test (i.e., the correlation between the item and the total score on the exam calculated excluding that item)." (*Id.*). A low item-total correlation means that an item has little relationship with the overall test and does not discriminate between those who perform well and those who perform poorly. Dr. Brink's report stated that items with "low item-total correlations should be dropped . . . because they are not operating in the intended manner and do not improve reliability." (*Id.*). Nunnally and Bernstein wrote an article concluding that items with an item-total correlation below .05 is "a very poorly discriminating item" and that items with an item-total correlation less than .2 "are at least moderately discriminating." (*Id.* at 42). Dr. Brink found that 45 items on the captain exam and 46 items on the senior-captain exam had item-total correlations below .2. He also found that 6 items on the captain exam had negative item-total correlation, indicating that high-scoring candidates were more likely to get the item incorrect than were low-scoring candidates.

Finally, Dr. Brink faulted the exams for producing statistically significant performance differences between black and white examinees. Dr. Brink found that 32 items on the captain exam and12 items on the senior-captain exam showed statistically significant differences based on a chi-square analysis using *p*-values of less than .05. (*Id.*). Dr. Brink argued that "[a]lthough changes to tests should not be made based solely on significant group differences, these items should be [the] focus of further evaluation to ensure that they are functioning appropriately." (*Id.*). The City had conducted no such evaluation.

Dr. Brink also calculated item bias through differential-item functioning ("DIF"). The SIOP Standards state that test developers should attempt to detect and eliminate aspects of test design, content, and format that may bias test scores for particular groups. DIF is intended to measure such sources of bias. Dr. Brink used the Mantel-Haenszel method to examine DIF. He stated that "[r]ace groups may differ with respect to performance on a particular item due to true differences (i.e, for some reason, there are real job-related differences between Blacks and Whites with respect to performance on the item within the sample) or race bias (i.e., there are not real job-related differences between Blacks and Whites with respect to performance on one item; for some reason, performance differences are occurring on the item because the item is biased against one of the race groups)." (*Id.* at 43). Dr. Brink's report stated that the Mantel-Haenszel analysis determines the extent to which black-white test-result differences are due to bias by examining the extent to which such differences exist after taking into account overall test performance. Relying on an article by Biddle, Dr. Brink stated that items with DIFs with *p*-values below .1 should be considered "meaningful" and those below .05 should be considered "even more substantial," but that items at both levels should be considered for removal from the test. (*Id.*). Dr. Brink found that 16 questions on the captain exam and 10 questions on the senior-captain exam showed bias. He believed that the City should have conducted DIF analysis and considered dropping these questions.

During the evidentiary hearing, Dr. Brink was asked about his findings on the multiple-choice examination format. Dr. Brink admitted that "to some degree," multiple-choice questions can measure more than job knowledge. (Evidentiary Hr'g Tr. 211, Docket Entry No. 130). But he also testified that skills such as communication and "interpersonal type of abilities" are poorly measured through such job-knowledge tests. (*Id.* at 212). Dr. Brink also testified that while written tests can measure leadership, command presence, and decision-making ability to a degree, there are

better ways to measure these skills and abilities. (*Id.* at 221). He also testified that situational-judgment questions better measure these skills and abilities than written job-knowledge questions, though he preferred "high-fidelity" exercises such as those performed at an assessment center over any type of written questions. (*Id.* at 222).

Dr. Brink also cited empirical research by Dr. Arthur. Dr. Brink testified that this research showed that "written tests with open-ended responses [were] actually more reliable than a written test with the closed-ended multiple-choice type responses." (*Id.* at 223). Dr. Arthur's study used a criterion-related validity study to compare the reliability of a multiple-choice exam to a "constructed response exam" that required the individuals to generate—rather than select—responses to exam questions. The construct-response questions were short-answer questions with a structured response format scored according to preestablished criteria. Winfred Arthur, Jr. *et al.*, *Multiple-Choice and Constructed Response Tests of Ability: Race-Based Subgroup Performance Differences on Alternative Paper-and-Pencil Test Formats*, 55 PERSONNEL PSYCHOLOGY 985, 996 (2002). The study found that the construct-response questions had higher reliability measures than multiple-choice questions. *Id.* at 998, 1000. The study also found that there was less subgroup difference from the construct-response questions, though the authors acknowledged that the sample size was small. *Id.* at 1001–02, 1004.

5. Dr. Lundquist's Affidavit and Testimony

Dr. Lundquist, a witness produced by the City, provided an affidavit and testimony on the validity of multiple-choice exams. She also testified about the validity of assessment centers. Dr. Lundquist concluded that the City's exclusive use of multiple-choice job-knowledge questions should be abandoned in favor of a promotional exam system incorporating assessment centers.

Pointing to the Guidelines, Dr. Lundquist stated that "[t]he emphasis for any promotional process should be on assessing the critical knowledge, skills, abilities, and other personal characteristics (KSAOs) identified through a job analysis as being required to perform the essential duties of the job." (Dr. Lundquist Aff. 3, Docket Entry No. 93-2). She acknowledged that the City's multiple-choice test could validly assess "the technical knowledge" required for the captain and senior-captain positions, but argued that such a test "inadequately captures the range of KSAOs required for successful performance in a position such as Senior Captain." (*Id.* at 4). Specifically, she argued that a multiple-choice test fails to test "supervisory and leadership skills and abilities." (*Id.*).

Dr. Lundquist's affidavit stated that situational-judgment questions and assessment centers better measure many of the abilities senior captains need. Dr. Lundquist argued that the literature shows that situational-judgment questions assess leadership and supervisory skills and abilities and should supplement, not replace, the multiple-choice job knowledge questions. Dr. Lundquist pointed to journal articles supporting the fairness and validity of assessment centers, as well as their ability to minimize disparate impact. During the evidentiary hearing, Dr. Lundquist discussed an article by Dr. Arthur concluding that assessment centers can validly measure "organization and planning and problem solving, . . . [and] influencing others." (Evidentiary Hr'g Tr. 233, Docket Entry No. 130). Dr. Arthur's study used "meta-analysis to empirically assess the criterion-related validity of separate dimensions tapped by assessment centers." Winfred Arthur, Jr. *et al.*, *A Meta-Analysis of the Criterion-Related Validity of Assessment Center Dimensions,* 56 PERSONNEL PSYCHOLOGY 125, 128 (2003). The study found "true validities" for the following performance dimensions: problem-solving, influencing others, and organizing and planning. *Id.* at 140.

Dr. Lundquist was asked about the objection that assessment centers produce "subjective" scores. She testified that through scoring standards and effective assessor training, assessment centers can produce scores approximating the objectivity of multiple-choice tests. (Evidentiary Hr'g Tr. 260–61, Docket Entry No 130). Dr. Lundquist admitted that scoring an assessment-center exercise is more subjective than scoring a multiple-choice test. But she argued that "reliability and consistency can be produced by certain controls in the design of the . . . assessment center exercise itself." (*Id.*). She also explained that objectivity and subjectivity are best understood as existing on a continuum, and that subjective scoring becomes more "objective" by structuring the scoring to minimize an assessor's subjective evaluation of a promotional candidate's performance. (*Id.* at 262–63). Dr. Lundquist testified that providing assessors with "very detailed examples of what is high performance, what is average performance, [and] what is low performance" for a particular performance dimension minimizes the assessor's subjective evaluation. (*Id.* at 264–65). She also testified that using multiple assessors reduces subjectivity. (*Id.* at 265).

On cross-examination, Dr. Lundquist admitted that "supervisory skill and planning and coordination skills" are difficult to measure and noted that "[w]hether or not it's a good assessment depends entirely on how well-written the test is and how similar it is to the requirements of the job." (*Id.* at 268). She testified that other methods besides a written test might be better measures of such skills. (*Id.*).

In response to questions about the importance of measuring cognitive skills, Dr. Lundquist acknowledged that industrial-psychology literature shows that "cognitive ability . . . underlies a lot of the performance, a lot of the learning that goes on in terms of any kind of job." (*Id.* at 235). But she emphasized that there are different types of cognitive skills and not all are important for the

captain and senior-captain positions. She explained that "you can think of cognitive ability as akin to an IQ test . . . . But if I think about a particular job, I don't want to know your IQ score. I want to know how well you do planning and decision making." (*Id.*). She testified that an assessment center would test the specific cognitive skills and abilities related to the captain and senior-captain jobs. (*Id.* at 236). For example, to test supervisory skills, "there will be some level of cognitive ability involved." (*Id.*). Dr. Lundquist, citing empirical studies, argued that "cognitively loaded" tests tend to produce subgroup differences. (*Id.* at 271). Dr. Lundquist testified that situational-judgment questions by themselves do not sufficiently measure such noncognitive skills and abilities as command presence, interpersonal communication, and leadership, because the questions are "cognitively loaded." Dr. Lundquist explained that situational-judgment questions require "a lot of reading." (*Id.* at 256). While asking such questions is "intended to measure application of knowledge," it "just does it in a way that is fairly complex in terms of getting to the question that people are having to answer." (*Id.*). She testified that "if you're trying to measure something that is not essentially cognitive, like interpersonal skills, . . . and you do that by requiring somebody to do a lot of reading, you'll get a cognitive component to the way the person performs on that test that's unrelated to what you're really trying to measure in the first place, which is interpersonal skills." (*Id.*). Dr. Lundquist testified that the situational-judgment test used for the November 2010 captain exam was "not that different from the job knowledge test" because "[i]t's another written multiple-choice test and even though you're calling it something different and trying to measure something more applied, it's not different enough to be producing a result that covers more of the space of the skills required." (*Id.* at 258).

6. Dr. David Morris's Testimony

Dr. Morris testified about the job analysis he had created for the captain position and about the job analysis he was creating for the senior-captain position. Dr. Morris also testified about the validity of the City's captain and senior-captain exams. Based on his work creating the job analyses, Dr. Morris concluded that the captain and senior-captain exams were not valid and offered alternatives to the City's promotional system.

To perform the job analysis for the captain position, Dr. Morris relied on subject-matter experts—both internal to the HFD and external to it—as well as source material. Dr. Morris's job analysis contained a more detailed listing of knowledge, skills, and abilities for the captain position than did the City's job description. Most of the identified "knowledges" related to: (1) firefighting, such as knowledge about equipment, structures, fires and firefighting and rescue tactics; (2) HFD standard operating procedures ("SOPs") and administrative processes; and (3) supervising. (Evidentiary Hr'g Ex. 42, at 2–6). The identified skills included those related to firefighting and operating firefighting equipment; leadership and communication; and problem-solving and decision-making. (*Id.* at 7). The identified abilities included lengthy lists of: (1) leadership abilities, including directing subordinates, resolving conflict, and motivating subordinates; (2) decision-making and strategic abilities such as prioritizing and developing contingency plans; (3) communication abilities, ranging from communicating with superiors to recognizing grammatical errors; (4) critical-thinking abilities, such as comprehending "complex rules, regulations, and procedures" and recognizing "critical aspects" of a problem; (5) administrative abilities, such as recording and documenting information; and (6) tactical abilities related to firefighting. (*Id.* at 8–11). Though Dr. Morris had not completed the job analysis for the senior-captain position, he testified that there was significant overlap with the captain position. He also testified that the senior-

captain position involved a "strong supervisory element" with a greater "span of control" than the captain position. (Evidentiary Hr'g Tr. 46–47, Docket Entry No. 130).

Based on the job analysis, Dr. Morris testified that the multiple-choice exam the City used does not reliably assess whether a promotional candidate is qualified for the captain or senior-captain position. Dr. Morris testified that while multiple-choice questions can be effective to test job knowledge, they have limited value in evaluating such skills and abilities as "communication, problem identification, . . . interpersonal skills, decision-making, and so forth," and these are the skills and abilities captains and senior captains should have. (*Id.* at 45). Dr. Morris also identified oral communication, command presence, and supervisory or interpersonal skills as skills poorly tested by a multiple-choice exam. (*Id.* at 48). Dr. Morris testified that the promotion exam measured only "a very small portion of the job—I wouldn't say a very small. It measures an important part of the job. But, in fact, the other part that is not measured is, in fact, very important for promotional and supervisory positions." (*Id.* at 66). On cross-examination, Dr. Morris acknowledged that a written test could measure more than knowledge and some KSAOs, and could measure problem-solving skills, but he emphasized that such a test would be unlikely to measure supervisory skills or oral-communication skills. (*Id*. at 83–85).

Dr. Morris testified that a promotional system incorporating a job-knowledge test and an assessment center would be valid. Dr. Morris did not initially recommend including a situational-judgment test but was comfortable with including it. (*Id.* at 45). Based on his job analysis, Dr. Morris concluded that there were "technical knowledge" requirements for the captain position that could be measured through a job-knowledge test. He testified that a situational-judgment question might also measure technical knowledge, but that "[i]t might be more likely that we would use a

knowledge-based test." (*Id.*). Dr. Morris admitted that "good practice" is usually to create a job analysis first, and then design a promotional test. (*Id.* at 78–79). He testified, however, that his experience helps him make accurate predications about what type of promotional test will be more reliable. (*Id.* at 77–78).

Dr. Morris's testimony emphasized the need for an assessment center, particularly for senior-captain promotions. He testified that such skills and abilities as supervision and leadership are best measured in assessment centers. Because senior captains have a great "span of control"—they directly supervise a large number of firefighters—an assessment center is uniquely reliable for evaluating senior-captain candidates. (*Id.* at 47–49). Dr. Morris also testified that an assessment center would be appropriate for captain promotion decisions. Though senior captains exercise a greater span of control than captains, similar skills and abilities are required for both positions. Dr. Morris described the captain position as a "gatekeeper" position for senior captain and argued that this justified examining similar skills and abilities. (*Id.* at 63).

At an assessment center, SMEs would create "standards of performance," tested by asking the candidates to respond to simulations of situations typically faced by captains and senior captains. The candidates would be scored by "assessors" according to preestablished performance standards, which would include examples of superior and inferior performance. (*Id.* at 68). Dr. Morris testified that this type of evaluation better predicts future job success because "it covers . . . a broader scope of the job and it's more accurate in its prediction." (*Id.* at 68). He also testified that,

assuming the center has assessors who score consistently, it allows the City to pick better captains and senior captains than it would pick through a multiple-choice test. (*Id.* at 70).[22]

Dr. Morris was also questioned about whether the proposed changes to the promotional system effectively measured cognitive abilities. Dr. Morris admitted that the proposed test changes would be designed to measure a candidate's ability to perform the job on the first day, not to test how the candidate would progress in the job. (*Id.* at 50). He testified that the Guidelines emphasize measuring a candidate's ability to start a job rather than to progress in it, based in part on the assumption that the candidate will receive additional training. (*Id.* at 52). But Dr. Morris disputed the contention that the proposed changes did not measure cognitive skills. He testified that cognitive abilities can be measured in ways other than a multiple-choice exam, such as through "structured interviews" and "structured oral exercises" that require exercising cognitive abilities to respond to questions and hypothetical situations. (*Id.* at 88).

7. Dr. Arthur's Testimony

Though Dr. Arthur did not submit an expert report, he testified on behalf of the HPFFA about the validity of many of the proposed changes to the captain and senior-captain exams. Dr. Arthur agreed with other experts' testimony that a job analysis should be the basis for developing an examination procedure. "[T]he method of assessment is informed by the job analysis process." (Evidentiary Hr'g Tr. 363, Docket Entry No 131). Dr. Arthur was skeptical of the validity of the changes proposed because the job analysis for the captain position was incomplete. He also noted that the proposals did not specify the changes in detail. (*Id.* at 365).

[22] Dr. Morris acknowledged that it could take significant time to develop an assessment center. (Evidentiary Hr'g Tr. 61–62, Docket Entry No 130). For example, producing video scenarios is a lengthy process that "might take a small Warner Bro[thers] production." (*Id.* at 61).

As to the use of multiple-choice job-knowledge questions, Dr. Arthur echoed the other experts' testimony that such tests can measure more than job knowledge, though he believed that certain skills and abilities are more effectively measured by other types of examinations. Dr. Arthur also pointed out that multiple-choice tests tend to be "cognitively loaded," but he argued that the issue is not the method of examination, rather the examination's "constructs," or content. (*Id.* at 365–66). He acknowledged that cognitively-loaded exams tend to create subgroup differences but argued that the reason for cognitive loading is not because the exam uses a multiple-choice test method. He explained that "assessment centers are likely to reduce subgroup differences not because you're using assessment centers, but because of the things that assessment centers measure." (*Id.* at 366; *see also id.* at 444). His criticism of this aspect of the proposed consent decree was that it focused on method and not content.

Dr. Arthur's testimony also provided a rough estimate based on the results from the 2010 captain exam as to whether there would be disparate impact in the number of promotional candidates eventually promoted. Dr. Arthur found that based on historical data, approximately 115 candidates are promoted each cycle. Using the rank order from the 2010 exam, Dr. Arthur found that there would not be a disparate impact under the 4/5 Rule if 110, 115, or 120 individuals were promoted to captain. (*Id.* at 383).

Dr. Arthur also testified about the proposed changes to the scoring system under the consent decree. For a valid examination method, "the relationship between test scores and performance is by definition linear in most situations." (*Id.* at 375). Dr. Arthur testified that banding tends to make performance differences obscure, contravening the "cardinal principle" of assessment. The better way to ensure score reliability is by ensuring the test is valid to begin with, so that "the same people

would get the same scores on the test no matter how many times they do it." (*Id.* at 376). Dr. Arthur was particularly critical of the pass/fail grading of the job-knowledge component. He acknowledged that such grading could be appropriate if only "minimal" job knowledge was required for the captain and senior-captain positions, but he had not seen evidence showing this to be true. (*Id.* at 416–17, 422–25).

Dr. Arthur also testified that the way in which a question is presented and the type of response the question demands can affect whether the exam produces subgroup differences. Asking a question in a format that requires reading and having the answer in a format requiring a written response cognitively load the question; examining the same content by presenting the question in a video format can reduce cognitive loading. (*Id.* at 424–27). Dr. Arthur pointed to an article he authored finding that asking questions requiring promotional candidates to generate, rather than select, answers can minimize subgroup differences. (*Id.* at 386–87).

8. William Barry's Testimony

William Barry, an Assistant Chief in the Houston Fire Department, testified on the City's behalf that "the captains who scored high on the multiple-choice tests were not always as effective as captains who scored lower on these tests. In fact, I never found a direct correlation between the scores on the test and their performance." (Evidentiary Hr'g Tr. 120, Docket Entry No 130). Barry, who currently works in HFD's "Member Support" — human resources — Division, explained that "taking tests and working in the four different ranks in the emergency operations and observing people who have been promoted through the system, the ability to memorize the correct 100 facts that are going to be asked on a test does not correlate to how these people perform in either complex personnel situations or emergency situations." (*Id.* at 121). Barry acknowledged that he has also

seen firefighters who did well on the exams perform well after promotion, though he thought the other direction was more common. (*Id.* at 134). He did admit that those who put the most time, energy, effort, and sacrifice into test preparation scored higher. (*Id.* at 137).

C. Summary

As noted above, the array of expert opinions about the existing exams' disparate impact on African-American candidates and the reliability and validity of the existing and proposed exams reflect the limits of the science of testing to measure and compare promotion-worthiness. The witnesses' opinions are at best attempts to gauge how well different exam approaches measure, compare, and predict job performance. But the absence of judicial expertise in the area of testing validity is well-recognized. The combination of the lack of judicial expertise in this area and the limits of the expertise of those who have training and experience support a cautious and careful judicial approach. With this background and caution, this court applies the legal standards to the evidence and opinions presented in the record.

III. Analysis

A. The Proposed Consent Decree and the HFPPA's Objections

1. The Standard of Review

"The parties to litigation may by compromise and settlement not only save the time, expense, and psychological toll but also avert the inevitable risk of litigation." *United States v. City of Miami*, 664 F.2d 435, 439 (5th Cir. 1981) (en banc) (Rubin, J., concurring). "Litigants . . . have sought to reinforce their compromise and to obtain its more ready enforceability by incorporating it into a proposed consent decree and seeking to have the court enter this decree." *Id.* "A consent decree, although founded on the agreement of the parties, is a judgment. It has the force of res judicata,

protecting the parties from future litigation. It thus has greater finality than a compact. As a judgment, it may be enforced by judicial sanctions, including citation for contempt if it is violated." *Id.* at 439–40 (internal citation and footnotes omitted).

"[A] decree disposing of some of the issues between some of the parties may be based on the consent of the parties who are affected by it but . . . to the extent the decree affects other parties or other issues, its validity must be tested by the same standards that are applicable in any other adversary proceeding."[23] *Id.* at 436. In discrimination suits that have produced proposed consent decrees that violate CBAs, the Fifth Circuit has held that "a court may not allow the substitution of a solution for past discrimination negotiated between the employer and the plaintiffs for that achieved through collective bargaining unless it first determines that the collectively bargained solution either violates Title VII or is inadequate in some particular to cure the effects of past discrimination." *Myers v. Gilman Paper Corp.*, 544 F.2d 837, 858 (5th Cir. 1977); *see also City of Miami*, 664 F.2d at 446–47 ("The right to promotion on the basis of test accomplishment may not be obliterated without a demonstration that the City has, in making promotions, discriminated against members of the affected classes in the past and that affirmative action is a necessary or appropriate remedy or that it has so discriminated in employment policy as to unfairly prejudice the

[23] The City and the plaintiffs argue that the "strong basis in evidence" standard announced in *Ricci v. DeStefano*, 129 S. Ct. 2658 (2009), provides the standard this court should use to evaluate the HPFFA's objections to the consent decree. In *Ricci*, seventeen white firefighters and one Hispanic firefighter sued the City of New Haven after it refused to certify results from a promotional examination after preliminary results showed disparate impact. The firefighters argued that the refusal to certify the results was an adverse employment decision on the basis of race. *Id.* at 2671. The issue in *Ricci* was whether "the purpose to avoid disparate-impact liability excuses what otherwise would be prohibited disparate-treatment discrimination." *Id.* at 2674. Here, the issue is not whether the City may take an adverse employment action against white firefighters by approving a promotional system that intentionally favors black candidates over white candidates. The issue is whether this court can approve a consent decree that violates state law and a collective-bargaining agreement based on the evidentiary record the parties have presented. *Ricci*'s "strong basis in evidence standard" is not applicable. Whether the Fifth Circuit precedent or *Ricci* applies, however, does not affect the outcome.

opportunity of the affected class to achieve promotions."); *League of United Latin Am. Citizens v. Clements*, 999 F.2d 831, 846 (5th Cir. 1993) (en banc) ("[O]ur preferences for settlement and accord are insufficient to justify the imposition of a decree that infringes upon the rights of third parties.").

The Fifth Circuit has advised district courts that:

> When presented with a proposed consent decree, the court's duty is akin, but not identical to its responsibility in approving settlements of class actions, stockholders' derivative suits, and proposed compromises of claims in bankruptcy. In these situations, the requisite court approval is merely the ratification of a compromise. The court must ascertain only that the settlement is "fair, adequate and reasonable."
>
> Because the consent decree does not merely validate a compromise but, by virtue of its injunctive provisions, reaches into the future and has continuing effect, its terms require more careful scrutiny. Even when it affects only the parties, the court should, therefore, examine it carefully to ascertain not only that it is a fair settlement but also that it does not put the court's sanction on and power behind a decree that violates Constitution, statute, or jurisprudence. This requires a determination that the proposal represents a reasonable factual and legal determination based on the facts of record, whether established by evidence, affidavit, or stipulation. If the decree also affects third parties, the court must be satisfied that the effect on them is neither unreasonable nor proscribed.

City of Miami, 664 F.2d at 441 (footnotes omitted). A district court should play a particularly "active role" when the litigation and settlement were instigated by a class of private plaintiffs—as opposed to the United States—because private plaintiffs have no "responsibility toward third parties who might be affected by their actions." *Williams v. City of New Orleans*, 729 F.2d 1554, 1560 (5th Cir. 1984) (en banc).

The threshold issue in the analysis is whether, under Title VII, the City's present testing method for senior-captain promotions has a disparate impact on African-American applicants.

B. Title VII

The plaintiffs and the City argue that the senior-captain exam disparately impacted promotions from the rank of captain to senior captain in violation of Title VII. "Title VII . . . prohibits employment discrimination on the basis of race, color, religion, sex, or national origin. Title VII prohibits both intentional discrimination (known as 'disparate treatment') as well as, in some cases, practices that are not intended to discriminate but in fact have a disproportionately disparate effect on minorities (known as 'disparate impact')." *Ricci v. DeStefano*, 129 S. Ct. 2658, 2672 (2009). "As enacted in 1964, Title VII's principal nondiscrimination provision held employers liable only for disparate treatment." *Id.* But in *Griggs v. Duke Power Co.*, 401 U.S. 424, 431 (1971), the Supreme Court "interpreted the Act to prohibit, in some cases, employers' facially neutral practices that, in fact, are 'discriminatory in operation.'" *Id.* at 2672–73. "Twenty years after *Griggs*, the Civil Rights Act of 1991, 105 Stat. 1071, was enacted. The Act included a provision codifying the prohibition on disparate-impact discrimination." *Id.* at 2673.

A plaintiff can make a *prima facie* case of discrimination by showing that an employer uses "a particular employment practice"—in this case a promotional test—"that causes a disparate impact on the basis of race." 42 U.S.C. § 2000e-2(k)(1)(A)(i). "To establish a *prima facie* case of discrimination under a disparate-impact theory, a plaintiff must show: (1) an identifiable, facially neutral personnel policy or practice; (2) a disparate effect on members of a protected class; and (3) a causal connection between the two." *McClain v. Lukin Indus., Inc.*, 519 F.3d 264, 275 (5th Cir. 2008) (citing *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977, 994 (1988)). "Ordinarily, a plaintiff must demonstrate that each particular challenged employment practice causes a disparate impact. Yet, Title VII provides that 'if the complaining party can demonstrate to the court that the

elements of [the employer's] decision making process are not capable of separation for analysis, the decision making process may be analyzed as one employment practice.'" *Id.* at 276 (quoting 42 U.S.C. § 2000e-2(k)(1)(B)(i)). An employer "may rebut a prima facie case of disparate impact by demonstrating that a challenged practice is a business necessity." *Crawford v. U.S. Dep't of Homeland Security*, 245 F. App'x 369, 379 (5th Cir. Aug. 16, 2007) (citing *Pacheco v. Mineta*, 448 F.3d 783, 787 (5th Cir. 2006)).

1. Disparate Impact

Disparate impact requires "a specific practice or set of practices resulting in a significant disparity between the groups." *Johnson v. Uncle Ben's, Inc.*, 965 F.2d 1363, 1367 (5th Cir. 1992). To establish a *prima facie* case, plaintiffs "must engage in a 'systematic analysis' of the policy or practice," *Frank v. Xerox Corp.*, 347 F.3d 130, 135 (5th Cir. 2003) (quoting *Munoz v. Orr*, 200 F.3d 291, 299 (5th Cir. 2000)), and "establish causation by offering statistical evidence to show that the practice in question has resulted in prohibited discrimination," *Stout v. Baxter Healthcare Corp.*, 282 F.3d 856, 860 (5th Cir. 2002). "Ordinarily, a *prima facie* disparate impact case requires a showing of a substantial 'statistical disparity between protected and non-protected workers in regards to employment or promotion.'" *Id.* (citing *Munoz*, 200 F.3d at 299–300); *see also Ricci*, 129 S. Ct. at 2678 (noting that a *prima facie* showing of disparate impact is "essentially, a threshold showing of a significant statistical disparity"); *Herndon v. Coll. of Mainland*, No. G-06-0286, 2009 WL 367500, at *28 (S.D. Tex. Feb. 13, 2009) ("[A *prima facie* showing] generally requires 'evidence of . . . observed statistical disparities,' but may include anecdotal evidence." (citations omitted)). While the Supreme Court has "emphasized the useful role that statistical methods can have in Title VII cases," it has "not suggested that any particular number of 'standard deviations' can determine

whether a plaintiff has made out a prima facie case in the complex area of employment discrimination." *Watson*, 487 U.S. at 995 n.3.

2. Business Necessity

It is a defense to liability if the challenged practice is shown to be "job related for the position in question and consistent with business necessity." 42 U.S.C. § 2000e-2(k)(1)(A)(i). "'The touchstone' for determining whether a test or qualification meets Title VII's measure, . . . is not 'good intent or the absence of discriminatory intent'; it is 'business necessity.'" *Ricci*, 129 S. Ct. at 2697 (quoting *Griggs*, 401 U.S. at 431). "When an employment test 'select[s] applicants for hire or promotion in a racial pattern significantly different from the pool of applicants,' . . . the employer must demonstrate a 'manifest relationship' between test and job." *Id.* (quoting *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 425 (1975)); *see also Frazier v. Garrison I.S.D.*, 980 F.2d 1514, 1526 n.34 (5th Cir. 1993) (explaining that the 1991 amendments to Title VII made clear that it is the employer's burden to show that a challenged practice is job-related and consistent with business necessity). The Fifth Circuit has held that *Albemarle Paper* requires a showing by professionally accepted methods that a test is "predictive of or significantly correlated with important elements of work behavior that comprise or are relevant to the job or jobs for which candidates are being evaluated." *Bernard v. Gulf Oil Corp.*, 841 F.2d 547, 564 (5th Cir. 1988).[24] "Such a showing . . . does not necessarily mean the employer prevails: '[I]t remains open to the complaining party to show that other tests or selection devices, without a similarly undesirable racial effect, would also

[24] The Fifth Circuit has also defined this burden as one to show that a "policy has a *significant* relationship to a legitimate business purpose." *EEOC v. J.M. Huber Corp.*, 927 F.2d 1322, 1328 (5th Cir. 1991).

serve the employer's legitimate interest in efficient and trustworthy workmanship.'" *Ricci*, 129 S. Ct. at 2697 (quoting *Albemarle Paper*, 422 U.S. at 425).[25]

C. Discussion

1. Disparate Impact

As set out above, on February 8, 2006, the City of Houston administered the senior-captain exam to 221 candidates. Of these, 172 were Caucasian, 15 were African-American, 33 were Hispanic, and 1 was "other." Of the 212 candidates passing by scoring 70 or above, 166 were Caucasian, 13 were African-American, 32 were Hispanic, and 1 was "other." The City promoted 70 individuals based on the rank order of those passing the exam. Of the 70 promoted, 59 were Caucasian, 2 were African-American, 8 were Hispanic, and 1 was "other." (Dr. McPhail Report 5).

The analysis starts with the Guideline "rule of thumb" for disparate impact, the 4/5 Rule. The expert witnesses in this case agreed that the total selection process for senior captains violated the 4/5 Rule. Under the Guidelines, a 4/5 Rule violation provides some evidence of disparate impact. *See* 29 C.F.R. § 1607.4(D) ("A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than

[25] The EEOC allows an employer to show that a test is job-related and consistent with business necessity—"valid"—through "criterion-related validity studies, content validity studies or construct validity studies." 29 C.F.R. § 1607.5(A). "To demonstrate the content validity of a selection procedure, a user should show that the behavior(s) demonstrated in the selection procedure are a representative sample of the behavior(s) of the job in question or that the selection procedure provides a representative sample of the work product of the job." 29 C.F.R. § 1607.14(C)(4). "Criterion related validation is established when there is a significant positive correlation between comparative success on the test (the 'predictor') and comparative success on some measure of job performance. The degree of correlation between test scores and job performance is expressed by a correlation coefficient. The value of the correlation coefficient can range from + 1.0 (employees with the highest test scores always perform better on the job) to -1.0 (employees with the highest test scores always perform worse on the job). A coefficient of zero indicates that there is no correlation between test and job performance." *Bernard*, 841 F.2d at 564.

four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact."); *Guidelines Questions & Answers*, 44 Fed. Reg. at 11999 (noting that when there is a 4/5 Rule violation, "[t]here usually is adverse impact"). But a 4/5 Rule violation is not enough to find disparate impact. Courts treat the 4/5 Rule as "no more than a 'rule of thumb' to aid in determining whether an employment practice has a disparate impact." *Waisome v. Port Auth. Of N.Y. & N.J.*, 948 F.2d 1370, 1376 (2d Cir. 1991); *see also Clady v. Cnty. of Los Angeles*, 770 F.2d 1421, 1428 (9th Cir. 1985) ("The Uniform Guidelines are not legally binding. They have not been promulgated as regulations and do not have the force of law." (citations omitted)); *Deshields v. Baltimore City Fire Dep't*, No. 88-3152, 1989 WL 100664, at *1 (4th Cir. Aug. 24, 1989) (unpublished) ("A rule of thumb used by a federal agency is not binding as a rule of law upon a federal court."); *Stagi v. Nat'l R.R. Passenger Corp.*, 391 F. App'x 133, 138 (3d Cir. Aug. 16, 2010) ("The '80 percent rule' or the 'four-fifths rule' has come under substantial criticism, and has not been particularly persuasive, at least as a prerequisite for making out a prima facie disparate impact case."). The Guidelines recognize that a 4/5 Rule violation does not mandate a disparate-impact finding, particularly when the population size of one group is small. *See* 29 C.F.R. § 1607.4(D) ("Greater differences in selection rate may not constitute adverse impact where the differences are based on small numbers and are not statistically significant . . . ."); *Guidelines Questions & Answers*, 44 Fed. Reg. at 11999 (stating that when the 4/5 Rule shows adverse impact, "[t]here usually is adverse impact, except where the number of persons selected and the difference in selection rates are very small). Similarly, the experts agreed that the 4/5 Rule does not always provide a reliable basis for finding disparate impact because it carries significant risks of statistical errors, particularly where sample sizes are small. (*See* Dr. McPhail Report 6–7; Dr. Morris Report 10446–47). This

conclusion is supported by literature in the field. S*ee* Boardman, *supra*, 25 at 776; Greenberg, *supra*, at 765–66; Morris & Lobsenz, *supra*, at 106 (concluding that the 4/5 Rule "does not take sampling error into account, and therefore, will often erroneously indicate adverse impact when none exists in the population"); Roth et al., *supra*, at 519–21.

Courts have recognized that when a group's population is small, statistical tests should be used to determine whether a 4/5 Rule violation is the product of chance. *See Black v. City of Akron*, 831 F.2d 131, 134 (6th Cir. 1987) ("[P]laintiffs are correct that the 4/5 rule has been criticized when used in the context of a small sample of employees being tested."); *Fudge v. City of Providence Fire Dep't*, 766 F.2d 650, 658 (1st Cir. 1985) ("We think that in cases involving a narrow data base, the better approach is for the courts to require a showing that the disparity is statistically significant, or unlikely to have occurred by chance, applying basic statistical tests as the method of proof."). Even outside the context of small group sizes, courts have recognized that statistical analyses provide stronger evidence of disparate impact than do violations of the 4/5 Rule standing alone. *See Clady*, 770 F.2d at 1428 ("[T]he 80 percent rule has been sharply criticized by courts and commentators."); *Isabel v. City of Memphis*, 404 F.3d 404, 412–13 (6th Cir. 2005) ("[W]e are grateful for statistics beyond the four-fifths rule analysis because we prefer to look to the sum of statistical evidence to make a decision in these kinds of cases."); *Stagi*, 391 F. App'x at 138 ("The '80 percent rule' or the 'four-fifths rule' has come under substantial criticism, and has not been particularly persuasive, at least as a prerequisite for making out a prima facie disparate impact case. The Supreme Court has noted that 'this enforcement standard has been criticized on technical grounds and it has not provided more than a rule of thumb for the courts.'" (alterations omitted) (quoting *Watson*, 487 U.S. at 995 n.3)); *see also* 1 B. Lindemann & P. Grossman, EMPLOYMENT DISCRIMINATION LAW 130 (4th

ed. 2007) (noting that the 80 percent rule "is inherently less probative than standard deviation analysis"); E. Shoben, *Differential Pass-Fail Rates in Employment Testing: Statistical Proof Under Title VII*, 91 HARV. L. REV. 793, 806 (1978) (arguing that the "four-fifths rule should be abandoned altogether" and that "flaws in the four-fifths rule can be eliminated by replacing it with a test of . . . statistical significance").[26]

In this case, the expert witnesses agreed that because the African-American promotion applicants were a relatively small group, statistical tests should be used to determine whether the 4/5 Rule violation resulted from chance. (Dr. Brink Report 10; Dr. Lundquist Aff. 6; Dr. Arthur Aff. 2–3; Dr. McPhail Report 6–7; Dr. Morris Report 10446–47). The 4/5 Rule violation provides evidence of, but does not establish, disparate impact.

As the HPFFA points out, courts frequently accept a .05 probability "to rule out the possibility that the disparity occurred at random." *Stagi*, 391 F. App'x at 137–38; *see also Page v. U.S. Indus., Inc.*, 726 F.2d 1038, 1047 n.5 (5th Cir. 1984) (noting that in *Castaneda v. Partida*, 430 U.S. 482, 496–97 n.17 (1977), the Supreme Court's "guidance" was that "a disparity in the number of minority workers in upper and lower level jobs is statistically significant if the difference between the expected number of minority employees in higher level positions exceeds the actual number by more than two or three standard deviations"); *Palmer v. Shultz*, 815 F.2d 84, 92–96 (D.C. Cir. 1987) (noting that "statistical evidence meeting the .05 level of significance is certainly sufficient to

[26] Courts use statistical evidence in ways the Guidelines do not. The Guidelines recognize the availability of statistical analysis but do not "rely primarily" on statistical tests because "[w]here the sample of persons selected is not large, even a large real difference between groups is likely not to be confirmed by a test of statistical significance (at the usual .05 level of significance)." *Guidelines Questions & Answers*, 44 Fed. Reg. at 11999. The Guidelines justify their approach because using "the 4/5ths rule of thumb [i]s a practical and easy-to-administer measure of whether differences in selection rates are substantial" and because "[m]any decisions in day-to-day life are made without reliance upon a test of statistical significance." (*Id.*).

support an inference of discrimination" (citation, internal quotation marks, and alterations omitted)); *Waisome*, 948 F.2d at 1376 ("Social scientists consider a finding of two standard deviations significant, meaning there is about one chance in 20 that the explanation for a deviation could be random and the deviation must be accounted for by some factor other than chance." (citation omitted)). The three generally accepted statistical tests for determining whether a 4/5 Rule violation resulted from chance—the Fisher exact, the Pearson chi-square, and the Z_d tests—did not show that the 4/5 Rule violation was statistically significant. These three tests did not provide evidence that the 2006 senior-captain exam disparately impacted African-Americans. (Dr. Brink Report 15; Dr. Arthur Aff. 3; Dr. McPhail Report 8–9; Dr. Morris Report 10447–49). Dr. Brink stated that under the Z_{ir} test, the 4/5 Rule violation was statistically significant. (Dr. Brink Report 14). But the scholarly articles the experts discussed suggest that the Fisher exact test was a better measure of statistical significance than the Z_{ir} test. One peer-reviewed journal article explains that while the Z_{ir} test was "interesting and deserve[s] greater thought," further research is needed into the test's ability to evaluate disparate impact. *See* Roth *et al.*, *supra*, at 520. The Z_{ir} test's creators acknowledged that the Fisher exact test will "provide a more accurate evaluation of statistical significance" than the Z_{ir} test when the smallest expected value in the analysis is less than five. Morris & Lobsenz, *supra*, at 97. Using a formula provided by Morris and Lobsenz, Dr. McPhail calculated the smallest expected value in the analysis to be 4.75 and concluded that the Z_{ir} test was inappropriate. (Dr. McPhail Report 9).

The statistical evidence does not support the conclusion that the 4/5 Rule violation for the 2006 senior-captain exam is not the product of chance. But courts have properly cautioned that "[t]here is no minimum statistical threshold requiring a mandatory finding that a plaintiff has

demonstrated a violation of Title VII" and "[c]ourts should take a 'case-by-case approach' in judging the significance or substantiality of disparities, one that considers not only statistics but also all the surrounding facts and circumstances." *Waisome*, 948 F.2d at 1376; *see also Int'l Bhd. of Teamsters v. United States*, 431 U.S. 324, 340 (1977) (noting that statistics "come in infinite variety and . . . their usefulness depends on all of the surrounding facts and circumstances"). The "surrounding circumstances" of the 2006 senior-captain exam provide evidence of disparate impact. Historical data shows that the promotions of African-Americans from captain to senior captain have violated the 4/5 Rule for every promotional cycle since 1993. (Docket Entry No. 93-2, at 8–9, 13). In two of those cycles, no black captains were promoted out of a total of 29 African-American applicants. (*Id.* at 13).

Even without statistical data confirming that there was statistically significant disparate impact in each of these promotional cycles, the historical data substantially mitigates the risk that the 4/5 Rule violation from the 2006 examination resulted from chance, even if the probability that the 2006 4/5 Rule violation is outside the standards of deviation courts commonly accept. Dr. Lundquist's Mantel-Haenszel analysis confirms that the historical patterns of 4/5 Rule violations are statistically significant.

Dr. Arthur argued that the only relevant data is for the promotions from the 2006 senior-captain exam. But courts have looked to historical data to help assess whether a promotional procedure is the result of chance. *See United Air Lines, Inc. v. Evans*, 431 U.S. 553, 558 (1977) ("A discriminatory act which is not made the basis for a timely charge . . . may constitute relevant background evidence in a proceeding in which the status of a current practice is at issue . . . ."); *Commonwealth of Pa. v. Flaherty*, 983 F.2d 1267, 1271 (3d Cir. 1993) (looking to past evidence of

disparate impact). The Guidelines also require analysis of historical data to assess disparate impact from current selection procedures. *See* 29 C.F.R. § 1607.4(D) ("Where the user's evidence concerning the impact of a selection procedure indicates adverse impact but is based upon numbers which are too small to be reliable, evidence concerning the impact of the procedure over a longer period of time and/or evidence concerning the impact which the selection procedure had when used in the same manner in similar circumstances elsewhere may be considered in determining adverse impact."). The historical data showing statistically significant disparate impact and the 4/5 Rule violation together support a finding of disparate impact from the 2006 senior-captain exam.

It is also worth noting that the Fisher exact test and the Pearson chi-square showed a 15% and 10% chance respectively that the 4/5 Rule violation was the product of chance. While higher than the 5% chance courts commonly accept, these percentages are not drastically higher. Dr. Brink's point that the 5% test is not a rigid standard to be applied under all circumstances applies here. One court made a similar point in finding age discrimination in firing decisions over objections that the evidence was outside the commonly accepted standards of deviation:

> The 5 percent test is arbitrary; it is influenced by the fact that scholarly publishers have limited space and don't want to clog up their journals and books with statistical findings that have a substantial probability of being a product of chance rather than of some interesting underlying relation between the variables of concern. Litigation generally is not fussy about evidence; much eyewitness and other nonquantitative evidence is subject to significant possibility of error, yet no effort is made to exclude it if it doesn't satisfy some counterpart to the 5 percent significance test. A lower significance level may show that the correlation is spurious, but may also be a result of "noise" in the data or collinearity (correlation between independent variables, such as sex and weight); and such evidence, when corroborated by other evidence, need not be deemed worthless. Conversely, a high significance level may be a misleading artifact of the study's design; and there is always the risk

> that the party's statistical witness ran 20 regressions, one and only one of which supported the party's position and that was the only one presented, though, in the circumstances, it was a chance result with no actual evidentiary significance.

Kadas v. MCI Systemhouse Corp., 255 F.3d 359, 362 (7th Cir. 2001); *see also* (Dr. Lundquist Test., Evidentiary Hr'g Tr. 276, Docket Entry No. 130 ("Well, there is a whole field of literature that talks about where to appropriately set that standard, should it be .05, should it be .10, should it be .01. So, although oftentimes in court you hear .05, and that is a common convention, . . . there's a lot of literature that suggests that depending on what the particular decision is that you're making, it may be more appropriate to use a different *P* value as your critical value for deciding whether something is statistically significant.")).

The HPFFA points out that some courts have also looked to whether small changes in the number of applicants in the groups at issue change the disparate-impact evidence based on the 4/5 Rule. *See Deshields*, 1989 WL 100064, at *1 (discrediting a 4/5 Rule violation as evidence of disparate impact because "[a] change in the race of only a few of those promoted could make a significant mathematical difference in the outcome of the four-fifths rule calculations"); *Waisome*, 948 F.2d at 1376 (noting that "if two additional black candidates passed the written examination the disparity would no longer be of statistical importance" and finding that the plaintiffs failed to make a *prima facie* showing of disparate impact). The record evidence shows that if two additional black captains had been promoted to senior captain in 2006, there would not have been a 4/5 Rule violation. Aside from the cases the HPFFA cites, however, there is no identified basis to conclude that because only two more black captains needed to be promoted for the promotional system to comply with the 4/5 Rule, this court should not find disparate impact. Given the small number of

black firefighters in the HFD compared to the number of white firefighters, it makes sense that only a few additional black captains would need to be promoted for the City to achieve compliance with the 4/5 Rule. But the historical data shows that the minimum number of African-American captain promotions to avoid a 4/5 Rule violation has never been achieved in any promotion cycle. Application of the Guidelines' n-of-1 rule and Dr. Brink's testimony about the one-person rule further support this point.

This court finds that the City and the plaintiffs have shown that the 2006 senior-captain exam disparately impacted the promotion of black captains to senior captain.[27] Because the City and the plaintiffs have made a *prima facie* showing of discrimination, it is the HPFFA's burden to show that the 2006 captain and senior-captain exams are job-related and consistent with business necessity.

2. Validity

The HPFFA has offered two expert reports to validate the captain and senior-captain exams.[28] Dr. McPhail's report provides a criterion-related validity study of the 2006 captain exam. It does not analyze the senior-captain exam. Dr. Sharf's report argues that both the captain and senior-captain exams are valid under a validity-generalization validity study. This court finds that neither report establishes that the captain and senior-captain exams are job-related and consistent with business necessity.

---

[27] Although the January 2011 evidentiary hearing focused on whether the 2006 senior-captain exam disparately impacted the promotion of black captains to senior captain, the City and the plaintiffs have also shown that the 2006 captain exam disparately impacted the promotion of firefighters to captain. The record reveals that the disparate-impact evidence for the 2006 captain exam is stronger than the disparate-impact evidence for the 2006 senior-captain exam. As Dr. Brink demonstrated, there was a 4/5 Rule violation for the 2006 captain exam that was statistically significant under the four relevant tests—the Fisher exact, the Pearson chi-square, the Z_d, and the Z_{ir}. (Dr. Brink Report 14). The HPFFA did not argue that the 2006 captain exam did not disparately impact the promotion of black firefighters to captain.

[28] The City has not conducted its own validity study. (Evidentiary Hr'g Tr. 373, Jan. 26, 2011, Docket Entry No. 131).

Dr. McPhail used a criterion-related validity study to analyze the results of the 2006 captain exam. The EEOC recognizes a criterion-related validity study as one of three validity measures of a promotional exam's validity. 29 C.F.R. § 1607.5(A). Courts have looked to the EEOC Guidelines validation procedures to analyze business necessity. *See Banos*, 398 F.3d at 892 ("The City can show that its process is 'job related' by any one of three tests: criterion-related, content validity, or construct validity."); *EEOC v. Dial Corp.*, 469 F.3d 735, 743 (8th Cir. 2006) (looking to the EEOC criterion validity requirements to evaluate business necessity); *Isabel*, 404 F.3d at 413 (relying on the EEOC Guidelines to determine validity); *United States v. City of Garland*, No. 3:98-CV-0307-L, 2004 WL 741295, at *9 (W.D. Tex. Mar. 31, 2004) (discussing criterion-related validation, content validation, and construct validation). Both the Guidelines and the case law recognize that an employer need demonstrate validity through only one method. *Hearn v. City of Jackson*, 340 F. Supp. 2d 728, 736 (S.D. Miss. 2003) ("'Neither the case law nor the Uniform Guidelines purports to require that an employer must demonstrate validity using more than one method.'" (quoting *Williams v. Ford Motor Co.*, 187 F.3d 533, 544–45 (6th Cir. 1999))); 29 C.F.R. § 1607.5(A) ("For the purposes of satisfying these guidelines, users may rely upon criterion-related validity studies, content validity studies or construct validity studies . . . ."); 29 C.F.R. § 1607.14(C)(1)) ("Users choosing to validate a selection procedure by a content validity strategy should determine whether it is appropriate to conduct such a study in the particular employment context."); *see also Washington v. Davis*, 426 U.S. 229, 248 n.13 (1976) (stating that "[i]t appears beyond doubt by now that there is no single method for appropriately validating employment tests for their relationship to job performance," and that any of the three recognized basic methods of validation may be used).

Dr. McPhail's report does not establish criterion-related validity for the 2006 captain exam. Dr. McPhail concluded that his statistical analyses showed only "equivocal evidence of the predictive capability of the 2006 examination." (Docket Entry No. 37-1, at 60). Dr. McPhail found statistically significant correlations above .05, suggesting validity, when he measured the entire validity sample. The Guidelines state that "[g]enerally, a selection procedure is considered related to the criterion, for the purposes of these guidelines, when the relationship between performance on the procedure and performance on the criterion measure is statistically significant at the 0.05 level of significance, which means that it is sufficiently high as to have a probability of no more than one (1) in twenty (20) to have occurred by chance." 29 C.F.R. § 1607.14(B)(5). But Dr. McPhail's report stated that these correlations supported two inferences, only one of which suggested validity. Either "the captain promotional examination effectively taps the intended construct domain which results in the observation that those scoring higher on the examination tend to have higher performance" or "because promotional examination scores were used as a basis for promotion . . . scores should be correlated with captain performance because those at the formal Captain rank have a greater opportunity to acquire knowledge and skills integral to effective functioning." (Docket Entry No 37-1, at 52). The second inference does not support finding that the 2006 captain exam is valid. Dr. McPhail's report did not conclude that the data supported one inference better than the other. Dr. McPhail's finding that the overall validation sample showed statistically significant correlations between test performance and job performance does not validate the 2006 captain exam.

Dr. McPhail used multiple-regression analysis to account for the additional training that promoted captains received, but he found statistically significant correlations for only three of the nine performance dimensions he measured. (*Id.* at 60). Similarly, when Dr. McPhail divided the

validation sample into promotional candidates who were promoted and job candidates who were not promoted, he found statistically significant correlations for only three of the nine performance dimensions, and he found those correlations only among the promotional candidates who were not promoted. Showing that the captain test produces significant correlations within one subgroup does not demonstrate that "the selection procedure is predictive of or significantly correlated with important elements of job performance." 29 C.F.R. § 1607.5(B). These correlations do not validate the test.

Dr. Sharf's validity-generalization validity study is no more persuasive. Dr. Sharf's report contains no analysis comparing the actual questions on the captain and senior-captain exams to the extensive list of cognitive skills he identified as essential to the positions. Instead, Dr. Sharf's report extensively describes literature supporting cognitive tests. He then describes the City's test as a cognitive test and concludes that it is valid. The Guidelines explicitly reject this form of validity analysis in the following section:

> Unacceptable substitutes for evidence of validity. Under no circumstances will the general reputation of a test or other selection procedures, its author or its publisher, or casual reports of it's validity be accepted in lieu of evidence of validity. Specifically ruled out are: assumptions of validity based on a procedure's name or descriptive labels; all forms of promotional literature; data bearing on the frequency of a procedure's usage; testimonial statements and credentials of sellers, users, or consultants; and other nonempirical or anecdotal accounts of selection practices or selection outcomes.

29 C.F.R. § 1607.9(A).

Even assuming that cognitive skills are a valid prediction of some aspects of job performance, Dr. Sharf's testimony does not provide a basis to conclude that the captain and senior-captain exams reliably measure the cognitive skills he identifies as necessary for those positions.

As Dr. Brink's expert report showed, some of the exam questions at best measure a promotional candidate's ability to memorize what appear to be obscure facts.

There is also substantial evidence in the record rebutting the HPFFA's evidence that the captain and senior-captain exams are job-related. Most of the experts who submitted a report or testified stated that both the captain and senior-captain positions require skills and abilities poorly measured by a multiple-choice exam. The experts identified the following nonexclusive list of such skills and abilities: leadership; command presence; interpersonal communication; supervision; and decision-making. (*See* Dr. Morris Test., Evidentiary Hr'g Tr. 45–48 (testifying that while multiple-choice questions can be effective at testing job knowledge, they do not adequately assess supervisory skills, communication, problem identification, interpersonal skills, decision-making, and command presence, which are skills and abilities that captains and senior captains should have); Dr. Brink Test., Evidentiary Hr'g Tr. 211–12 (admitting that "to some degree" a written test can measure more than job knowledge, but emphasizing that skills such as communication and "interpersonal type of abilities" are poorly measured through written, multiple-choice job-knowledge tests); Dr. Lundquist Test., Evidentiary Hr'g Tr. 239 (stating that multiple-choice job-knowledge tests typically "cover . . . a more limited set of skills than might be required for [the captain and senior captain] positions"); Dr. Lundquist Aff. 4 (acknowledging that the City's multiple-choice test could validly assess the technical knowledge required for the positions of captain and senior captain, but arguing that such a test "inadequately captures the range of KSAOs required for successful performance in a position such as Senior Captain"); *see also* Dr. Lundquist Test., Evidentiary Hr'g Tr. 240 (explaining that "as you move up through supervisory ranks, . . . what you see is less of an emphasis on the technical or knowledge side and more of an emphasis on

. . . leadership, supervisory, managerial, [and] strategic . . . aspects of the job"); Dr. McPhail Validation Study of the 2006 Captain Exam, Docket Entry No. 37-1, at 37 (identifying "supervision," "problem solving," "interpersonal effectiveness," and "professional orientation & commitment," as comprising four of the nine performance dimensions for the captain position); Dr. Sharf Report 17–21 (listing management and supervision, problem solving, communication, command presence, and leadership as responsibilities of the captain and senior-captain positions)).

The evidence shows that the promotional tests for the captain and senior-captain positions did not test the entire "job domain." Courts have rejected promotional tests for similar reasons. *See Isabel*, 404 F.3d at 413 (upholding district court's determination that because the test only examined "job knowledge," it failed to test the entire "job domain"). The Guidelines similarly suggest that promotional examinations should replicate work behaviors. *Guidelines Questions & Answers*, 44 Fed. Reg. at 12007. The expert testimony demonstrates that the captain and senior-captain exams fail to test significant elements of the positions.[29]

---

[29] The Supreme Court has noted that "[i]t is self-evident that many jobs, for example those involving managerial responsibilities, require personal qualities that have never been considered amenable to standardized testing." *Watson*, 487 U.S. at 998; *see also Puffer v. Allstate Ins. Co.*, 255 F.R.D. 450, 459–60 (N.D. Ill. 2009) ("Indeed, it is hard to envision a system that could thoroughly evaluate the suitability of persons for management jobs, or their performance in those positions, without taking into account subjective assessments of qualities that objective criteria may be unable to capture."). Courts have also found that subjective criteria are particularly important for supervisory positions. The Fifth Circuit has stated that "[s]ubjective criteria necessarily and legitimately enter into personnel decisions involving supervisory positions." *Risher v. Aldridge*, 889 F.2d 592, 597 (5th Cir. 1989). Similarly, the Seventh Circuit has stated:

> Subjective evaluations of a job candidate are often critical to the decisionmaking process, and if anything, are becoming more so in our increasingly service-oriented economy. Personal qualities factor heavily into employment decisions concerning supervisory or professional positions. Traits such as common sense, good judgment, originality, ambition, loyalty, and tact often must be assessed primarily in a subjective fashion, yet they are essential to an individual's success in a supervisory or professional position.

Blise v. Antaramian, 409 F.3d 861, 868 (7th Cir. 2005) (alterations omitted) (quoting *Chapman v. A.I. Transport*, 229 F.3d 1012, 1033–34 (11th Cir. 2000)).

Dr. Brink's report also provides persuasive evidence that the captain exam itself does not produce reliable scores according to standards accepted by professional industrial psychologists. The Supreme Court has observed that "[t]he message of these Guidelines is the same as that of the *Griggs* case—that discriminatory tests are impermissible unless shown, by professionally acceptable methods, to be predictive of or significantly correlated with important elements of work behavior which comprise or are relevant to the job." *Albemarle Paper*, 422 U.S. at 431 (internal quotation marks omitted). Dr. Brink's item analysis of the 2006 captain and senior-captain exams showed that the scores did not reliably demonstrate superior knowledge even for the limited areas of knowledge related to the captain and senior-captain positions. Dr. Brink found that based on "item difficulty" analysis, 34 items on the captain exam and 60 items on the senior-captain exam should have been excluded. Based on "index discrimination analysis," Dr. Brink found that 53 items on the captain exam and 66 items on the senior-captain exam should have been excluded. Finally, based on "item-correlation analysis," 45 items on the captain exam and 46 items on the senior-captain exam should have been excluded. Dr. Brink testified that a reliable examination is prerequisite to a valid examination: "for an employment test to accurately predict job performance, it *must* be reliable; but having a reliable test does not guarantee accurate prediction of job performance." (Dr. Brink Report 39). Dr. Brink's testimony shows that the scoring of the captain and senior-captain exams is inconsistent with professionally accepted standards of reliability and provides additional evidence that the exams were not valid.

Dr. Brink's testimony also provided evidence that the captain and senior-captain exams did not measure the KSAOs identified in the captain and senior-captain job descriptions. The Guidelines require that "[f]or any selection procedure measuring a knowledge, skill, or ability the

user should show that (a) the selection procedure measures and is a representative sample of that knowledge, skill, or ability; and (b) that knowledge, skill, or ability is used in and is a necessary prerequisite to performance of critical or important work behavior(s)." 29 C.F.R. § 1607.14(C)(4). Dr. Brink's report noted that 63% of the captain exam content and 86% of the senior-captain exam content were not necessary for the applicant to perform the duties of the first day of work. This conclusion, based on his analysis of the exams and his consultations with SMEs, provides further evidence of the lack of validity for the exams.

The HPFFA has not demonstrated that the captain and senior-captain promotion exams are job-related and consistent with business necessity. Because this court has found disparate impact for both the captain and senior-captain exams, and because the HPFFA has failed to show that the exams are job-related and consistent with business necessity, the City may implement changes to the promotional system for the captain and senior-captain positions to the extent necessary to address the disparate impact, even if those changes are inconsistent with the CBA and state law.

The City and the plaintiffs have demonstrated that the captain and senior-captain exams disparately impacted black promotional candidates and that the promotional examinations for the positions are not justified by business necessity. This court's finding provides a basis to approve provisions of the proposed consent decree that conflict with the TLGC and the CBA. But this court must be cautious to approve only those conflicting provisions of the proposed consent decree that are necessary and tailored to remedy the demonstrated disparate impact.

IV. The Consent Decree

A. The Legal Standards

"[A]ny federal decree must be a tailored remedial response to illegality." *Clements*, 999 F.2d at 847 (citing *Shaw v. Reno*, 509 U.S. 630 (1993)). "A consent decree must arise from the pleaded case and further the objectives of the law upon which the complaint is based." *Id.* at 846; *see also San Antonio Hispanic Police Officers' Org. v. City of San Antonio*, 188 F.R.D. 433, 439 (W.D. Tex. 1999) ("[T]he question for the courts is whether this part of the proposal has a sufficient nexus to the lawsuit to justify circumventing the collective bargaining process."). "'[F]ederal-court decrees exceed appropriate limits if they are aimed at eliminating a condition that does not violate [federal law] or does not flow from such a violation.'" *Horne v. Flores*, 129 S. Ct. 2579, 2595 (2009) (quoting *Milliken v. Bradley*, 433 U.S. 267, 282 (1977)); *see also Milliken*, 433 U.S. at 281–82 ("The well-settled principle that the nature and scope of the remedy are to be determined by the violation means simply that federal-court decrees must directly address and relate to the constitutional violation itself."). "Courts must be exceptionally cautious when litigants seek to achieve by consent decree what they could not achieve by their own authority. Consent is not enough when parties seek to grant themselves powers they do not hold outside of court." *City of San Antonio*, 188 F.R.D. at 458 (citing *Clements*, 999 F.2d at 846).

Courts must be particularly cautious when a consent decree conflicts with a collective-bargaining agreement. "[P]arties to a collective-bargaining agreement must have reasonable assurance that their contract will be honored." *W.R. Grace & Co. v. Local Union 759, Int'l Union of United Rubber, Cork, Linoleum & Plastic Workers of Am.*, 461 U.S. 757, 771 (1983). "[R]egardless of past wrongs, a court in considering prospective relief is not automatically empowered to make wholesale changes in agreements negotiated by the employees' exclusive

bargaining agents in an obviously serious attempt to comply with Title VII." *Myers*, 544 F.2d at 857.

B. Discussion

The City and the plaintiffs have demonstrated that allowing the City to use situational-judgment questions and an assessment center to examine promotional candidates for the positions of captain and senior captain is tailored to remedy the disparate impact alleged in the plaintiffs' complaint and to ensure that the City's promotional processes for the positions are job-related and consistent with business necessity. The plaintiffs and the City have shown that incorporating situational-judgment tests and assessment centers diminishes the risk of disparate impact and increases the validity of the City's promotional processes. Dr. Brink credibly testified that situational-judgment tests better measure command presence, decision-making ability, and leadership than multiple choice tests. (Dr. Brink Test., Evidentiary Hr'g Tr. 221–22). There was also testimony that incorporating a situational-judgment component into the promotional examination process reduces the risk of disparate impact. Both Dr. Brink and Dr. Arthur testified that situational-judgment questions can be designed to minimize cognitive loading. Dr. Arthur testified that displaying a hypothetical situation by video instead of describing it in words reduces the amount of cognitive skill required to answer the situational-judgment question correctly. Dr. Brink, describing an article by Dr. Arthur, showed that asking questions that require promotional candidates to generate, rather than select, the correct answer can help reduce subgroup differences. (*Id.* at 223). Dr. Arthur's testimony that, based on the number of promotions made during past promotional cycles and the results from the 2010 captain exam incorporating situational-judgment questions, there may not be disparate impact for the 2010–2013 promotional cycle also provides

circumstantial evidence that the situational-judgment component of the exam may help reduce the risk of disparate impact.

But the evidence also showed that situational-judgment questions alone are unlikely to measure the entire job domain for the captain and senior-captain positions. Dr. Lundquist, Dr. Brink, and Dr. Arthur all testified that situational-judgment tests are not as effective as assessment centers at measuring skills and abilities the experts agreed were important, such as command presence, leadership, and interpersonal communication. (*See* Dr. Brink Test., Evidentiary Hr'g Tr. 221–22; Dr. Lundquist Test., Evidentiary Hr'g Tr. 233 (discussing an article by Dr. Arthur concluding that assessment centers can validly measure "organization and planning and problem solving, . . . [and] influencing others")). Dr. Lundquist also explained that situational-judgment questions inevitably involve some cognitive loading, which dilutes their ability to measure noncognitive skills. (*Id.* at 256). The experts, including Dr. Arthur, acknowledged that limiting the role of the assessment center to measuring the types of skills best measured by such centers also reduces the risk of disparate impact.

The HPFFA's argument that a promotional system should not be designed until job analyses are complete is not a sufficient basis to reject the proposed consent decree. Most of the experts agreed that both the captain and senior-captain positions require skills and abilities—including, but not limited to, command presence, supervision, leadership, interpersonal communication, and decision-making—that are poorly measured by a multiple-choice test. Dr. Lundquist and Dr. Brink, citing to empirical studies widely accepted by industrial psychologists, also showed that cognitively loaded multiple-choice tests further entrenched the disparate impact of a multiple-choice test format.

Dr. Arthur's distinction between the content of the tests and the multiple-choice format as factors in causing disparate impact is an important one. But as Dr. Arthur acknowledged, multiple-choice tests are generally used to analyze cognitively loaded content. The City's continued reliance on exclusively multiple-choice test formats risks continued disparate impact that is neither job-related nor justified by business necessity.

The HPFFA objects that using assessment centers risks subjective scoring. But this objection is not a basis for rejecting this part of the consent decree. There was credible evidence in the record that there are ways to make the assessment center's results less subjective. For example, Dr. Lundquist explained that by using scoring standards and effective assessor training, assessment centers can produce scores approximating the objectivity of multiple-choice tests. (Evidentiary Hr'g Tr. 260–61, Docket Entry No. 130). Establishing detailed, predetermined criteria for performance; providing thorough training to assessors; using multiple assessors; and using postexamination statistical analyses to account for subjective differences between individual assessors, are all recognized as effective ways to reduce the risk that promotional candidates will be scored based on subjective judgments unrelated to the quality of their performance in the assessment-center exercises.

While the City and the plaintiffs have demonstrated that the use of situational-judgment questions and an assessment center are both tailored to minimize disparate impact and to validate the City's promotional processes for captain and senior-captain positions, the City and the plaintiffs have not demonstrated that the proposed consent decree's wholesale abandonment of promotion based on superior performance on other components of competitive exams is needed either to reduce disparate impact or to validate the promotional processes. The scoring system the consent decree

proposes makes a promotional candidate's assessment-center score—which the evidence shows only validly measures part of the "job domain" for the captain and senior-captain positions—the most important score in the promotional process. The evidence did not show that the skills and abilities best measured by an assessment center are the only or most important skills for the captain and senior-captain positions.

The proposed consent decree allows either no cognitive substantive job-knowledge test or allows a scoring system that makes a job-knowledge test pass/fail, negating the value of superior performance on that test. Dr. Arthur's testimony that a pass/fail test would be appropriate if there were only minimal competency requirements for the captain and senior-captain positions is credible and persuasive. (*See* Evidentiary Hr'g Tr. 416–17, Docket Entry No 131 ("But I would say if you have some basis for wanting to use pass/fail, which is a minimal competency approach, then there ought to be some argument articulated as to why for this particular exam a minimal competency approach is appropriate, whereas it's not for others.")). The City and the plaintiffs have not demonstrated that "minimal knowledge" is sufficient for either position. To the contrary, the evidence and testimony was that both captain and senior-captain positions had significant knowledge requirements and there was no evidence or testimony that the knowledge required was minimal. (*See* Dr. McPhail Validation Study of the 2006 Captain Exam, Docket Entry No. 37-1, at 37 (identifying "technical knowledge" as one of nine performance dimensions for the captain position based on HFD job descriptions and interviews with SMEs); Dr. Lundquist Aff. 3–4, 8 (stating that multiple-choice tests adequately assess the technical knowledge required for the captain and senior-captain positions, but arguing that situational-judgment tests and assessment-center exercises should be added to the promotional exam to measure the other critical KSAOs); Dr. Morris Test.,

Evidentiary Hr'g Tr. 66 (noting that a job-knowledge test "measures an important part of the job")). Dr. McPhail's criterion-related validity study also demonstrated that the City's job-knowledge tests have produced statistically significant correlations between the KSAOs related to the station management, resources management, and problem-solving performance dimensions even after accounting for on-the-job training. (Docket Entry No. 37-1, at 60–61). The consent decree, however, would use a passing score on the knowledge test only to determine which candidates could proceed to the next steps of the exam. Although the record shows that exclusive use of a multiple-choice job-knowledge test is linked to disparate impact, the record does not show that reducing the job-knowledge test to one part of the promotion exam and grading it pass/fail—and making it otherwise irrelevant to the promotion decision—are steps tailored to the disparate impact remedy. *Cf. Clements*, 999 F.2d at 847 (stating that "any federal decree must be a tailored remedial response to illegality"); *United States v. City of New York*, 637 F. Supp. 2d 77, 118–21 (E.D.N.Y. 2009) (holding that a promotional exam was not job-related when it failed to test important cognitive and noncognitive abilities).

It appears that once a candidate's assessment-center scores are determined, the candidate's performance on either the job-knowledge test or the situational-judgment test are irrelevant to the promotion decision. Instead, the assessment-center scores are "banded" and promotional decision made based on those bands. The evidence is that in addition to the substantive knowledge measured by the job-knowledge component, the situational-judgment component provides reliable and valid measures of many of the skills and abilities relevant to the captain and senior captain "job domain." Under the proposed consent decree, these scores will only be of very limited value in determining whether a promotional candidate should be promoted. That is true for the situational-judgment-test

component even though the evidence shows that it could validly measure the KSAOs directly relevant to whether a promotional candidate will succeed as a captain or senior captain. The record does not justify these consent-decree provisions as tailored to remedy the disparate impact of the existing promotion exam process.

To the contrary, the evidence shows that a job-knowledge test measuring the knowledge needed for promotion and a situational-judgment test measuring other KSAOs needed for promotion will likely produce some reliable and valid measures of a promotional candidate's ability to perform. The evidence also showed that incorporating situational-judgment questions and an assessment center into the promotion process in addition to the job-knowledge tests would reduce disparate impact. The evidence showed that a promotional examination using a job-knowledge component, a computer-based situational-judgment component, and an assessment-center component would measure the KSAOs required for promotion with more reliability and validity than using only one or even two of the components and would reduce disparate impact. But there is insufficient evidence that giving the primacy to the assessment center as proposed in the modified consent decree is also needed to reduce adverse impact or to create a promotional system that is job-related and consistent with business necessity.

Similarly, there is not a sufficient evidentiary basis to abandon the Rule of Three codified in the TLGC and accepted in the CBA. Dr. Brink and Dr. Lundquist testified that statistical studies may show that for promotional candidates within a "band of scores," marginally better exam performance does not correlate to better job performance after promotion. Even accepting this testimony, however, the proposed consent decree replaces one clear and predictable method of selecting among promotional candidates with similar scores—the Rule of Three—with a decision-

maker's subjective judgment and discretion. The choice of one candidate within a band is standardless. Dr. Brink and Dr. Lundquist both testified that within a band, no one candidate is viewed as more qualified than another. Neither testified that the decision-maker's discretion will produce better selections than the Rule of Three. None of the experts testified that race was to be a criterion for selecting from within a band.

There is no evidence that replacing the Rule of Three is necessary to reduce disparate impact or that banding is likely to be a more valid or reliable basis for identifying job-related qualifications than the Rule of Three. To the contrary, there was evidence that if the test is valid and reliable, there is a "linear relationship" between test scores and performance. (Evidentiary Hr'g Tr. 375, Docket Entry No. 131); *accord Nash v. Consolidated City of Jacksonville, Duval Cnty., Fla*, 895 F. Supp. 1536, 1551 (M.D. Fla. 1995) (summarizing expert testimony showing that if a promotional test is valid and reliable, the test "served as a good predictor of success for the job"). The City has hired industrial psychology experts to design valid and reliable promotional exams. This evidence supports retaining the TLGC and CBA provisions relating to the Rule of Three, which requires the decision-maker to select the promotional candidate with the highest score unless there is a written explanation justifying a different choice. *See id.* at 1552 (upholding Jacksonville's "rule of one" where the promotional exam was "highly reliable and the City demonstrated the substantial job-relatedness of the exam"); *cf. id.* at 1553 (stating that eliminating Jacksonville's "rule of one" would "open the [promotional] process to favoritism, politics and tokenism, just what the City is trying to avoid by using the rule of one"). On the basis of the current record, the consent decree's provision abandoning the Rule of Three is not tailored to remedy disparate impact.

V. Conclusion

The proposed consent decree's provisions requiring a situational-judgment and an assessment-center component to the promotional exams are approved. Provision "1" of the proposed consent decree is also approved. Provision "4a" of the proposed consent decree, to which the HPFFA has not objected, and which requires the City to bargain for a different seniority system, is also approved.

The following provisions in the decree are not approved because they violate the TLGC and CBA and the City and the plaintiffs have not shown that they are tailored to remedy the statutory violation:

> 2. Job-Knowledge Written Test
> - Pass/Fail Exam (test designer to determine cut off score)
> - No rank-order list
> - Test designer may elect not to use a written job knowledge cognitive test
>
> 3. Scenario-Based Computer-Objective Test
> - Rank-order list from which all intended promotions to assessment center will be made
>
> 4. Assessment Center
> - Rank-order list
> - Sliding bands based on test accuracy as determined by consultant
> - Fire Chief will document reasons for selecting each candidate within bands
> - No Rule of Three

(Docket Entry No. 69-2, at 29).

Devising the proper method for scoring the examination should be accomplished by collective bargaining and after the completion of the job analysis for the senior-captain position. This is consistent both with best methods for test development and with the principles embodied in Title VII—which favors voluntary compliance—and the TLGC, CBA, and Fifth Circuit precedent.

A hearing is set for **February 21, at 1:30 p.m.** in Courtroom 11-B to address the issues that remain to be resolved and set a timetable for doing so.

SIGNED on February 6, 2012, at Houston, Texas.

Lee H. Rosenthal
United States District Judge